Springer Graduate Texts in Philosophy

Horacio Arló-Costa · Vincent F. Hendricks
Johan van Benthem *Editors*
Henrik Boensvang · Rasmus K. Rendsvig
Assistant Editors

Readings in Formal Epistemology

Sourcebook

Springer

Springer Graduate Texts in Philosophy

Volume 1

The Springer Graduate Texts in Philosophy offers a series of self-contained textbooks aimed towards the graduate level that covers all areas of philosophy ranging from classical philosophy to contemporary topics in the field. The texts will, in general, include teaching aids (such as exercises and summaries), and covers the range from graduate level introductions to advanced topics in the field. The publications in this series offer volumes with a broad overview of theory in core topics in field and volumes with comprehensive approaches to a single key topic in the field. Thus, the series offers publications for both general introductory courses as well as courses focused on a sub-discipline within philosophy.

The series publishes:

- All of the philosophical traditions
- Includes sourcebooks, lectures notes for advanced level courses, as well as textbooks covering specialized topics
- Interdisciplinary introductions – where philosophy overlaps with other scientific or practical areas

We aim to make a first decision within 1 month of submission. In case of a positive first decision the work will be provisionally contracted: the final decision about publication will depend upon the result of the anonymous peer review of the complete manuscript. We aim to have the complete work peer-reviewed within 3 months of submission. Proposals should include:

- A short synopsis of the work or the introduction chapter
- The proposed Table of Contents
- CV of the lead author(s)
- List of courses for possible course adoption.

The series discourages the submission of manuscripts that are below 65,000 words in length.

For inquiries and submissions of proposals, authors can contact Ties.Nijssen@Springer.com

More information about this series at http://www.springer.com/series/13799

Horacio Arló-Costa • Vincent F. Hendricks
Johan van Benthem
Editors

Readings in Formal Epistemology

Sourcebook

Assistant Editors
Henrik Boensvang
Rasmus K. Rendsvig

Springer

Editors
Horacio Arló-Costa (deceased)

Johan van Benthem
University of Amsterdam
Amsterdam, The Netherlands

Stanford University
Stanford, United States

Assistant Editors
Henrik Boensvang
University of Copenhagen
Copenhagen, Denmark

Vincent F. Hendricks
Center for Information and Bubble Studies
University of Copenhagen
Copenhagen, Denmark

Rasmus K. Rendsvig
Lund University
Lund, Sweden

Center for Information and Bubble Studies
University of Copenhagen
Copenhagen, Denmark

Springer Graduate Texts in Philosophy
ISBN 978-3-319-20450-5 ISBN 978-3-319-20451-2 (eBook)
DOI 10.1007/978-3-319-20451-2

Library of Congress Control Number: 2016935072

Springer Cham Heidelberg New York Dordrecht London
© Springer International Publishing Switzerland 2016
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

On June 1, 2011, we received an email from Horacio, expressing his pleasure in our successful cooperation that was drawing to a close and his eager anticipation of the published book. This would turn out to be the last correspondence we received from our friend. Horacio died on July 14, 2011. This anthology would not have seen the light of day without Horacio's encyclopedic knowledge and fine scholarship, his original angle of approach to philosophy, and his constant ambition to push the formal epistemology agenda forward. Horacio will never see our joint work in print, but we will continue his quest and honor his legacy by dedicating this book to him.

Vincent F. Hendricks
Johan van Benthem

Preface

"Formal epistemology" is a term coined in the late 1990s for a new constellation of interests in philosophy, merging traditional epistemological concerns with new influences from surrounding disciplines like linguistics, game theory, and computer science. Of course, this movement did not spring to life just then. Formal epistemological studies may be found in the classic works of Carnap, Hintikka, Levi, Lewis, Kripke, Putnam, Quine, and many others.

Formal epistemology addresses a growing agenda of problems concerning knowledge, belief, certainty, rationality, deliberation, decision, strategy, action, and agent interaction – and it does so using methods from logic, probability theory, computability theory, decision theory, game theory, and elsewhere. The use of these formal tools is to rigorously formulate, analyze, and sometimes solve important issues of interest to philosophers but also to researchers in other disciplines, from the natural sciences and humanities to the social and cognitive sciences and sometimes even the realm of technology. This makes formal epistemology an interdisciplinary endeavor practiced by philosophers, logicians, mathematicians, computer scientists, theoretical economists, social scientists, cognitive psychologists, etc.

Although a relative newcomer, formal epistemology is already establishing itself in research environments and university curricula. There are conferences, workshops, centers, and jobs in formal epistemology, and several institutions offer courses or seminars in the field.

Yet no volume is in existence comprising canonical texts that define the field by exemplars. Lecturers and students are forced to collect influential classics and seminal contemporary papers from uneven sources, some of them hard to obtain even for university libraries. There are excellent anthologies in mainstream epistemology, but these are not tuned to new fruitful interactions between the mainstream and a wider spectrum of formal approaches.

Readings in Formal Epistemology is intended to remedy this situation by presenting some three dozen key texts, divided into five subsections: Bayesian Epistemology, Belief Change, Decision Theory, Logics of Knowledge and Belief, and Interactive Epistemology. The selection made is by no means complete but

hopefully representative enough for an accurate picture of the landscape. This collection will hopefully serve as a study and research companion while also helping shape and stimulate a flourishing new field in philosophy and its broader intellectual environment.

Pittsburgh, PA, USA — Horacio Arló-Costa
Copenhagen, Denmark — Vincent F. Hendricks
Amsterdam, The Netherlands — Johan van Benthem
Copenhagen, Denmark — Henrik Boensvang
Lund, Sweden — Rasmus K. Rendsvig

Acknowledgments

On the way to compiling this volume, we have been assisted by many people. Jeffrey Helzner and Gregory Wheeler gave us valuable suggestions for texts to include. We are grateful for their assistance in the selection process. Many authors represented in this volume provided us with essential copies of their papers while also giving important input on the organization of this collection. We thank them for their kind help. We would have liked to have included even more seminal papers, but due to limitations of space, and the fact that some copyrights were either impossible trace or too expensive to obtain, we ended up with the current selection. We are furthermore indebted to Springer Science and Business Media for taking on this project, especially Ties Nijssen, Christi Lue, and Werner Hermens. The editors also acknowledge the generous funding provided by the Elite Research Prize from the Danish Ministry of Science, Technology, and Innovation awarded to Vincent F. Hendricks in 2008.

Finally, this volume would not have seen the light of day without the constant efforts of Henrik Boensvang and Rasmus K. Rendsvig in communicating with relevant parties, collecting the required permissions, and compiling all the papers patiently and efficiently while paying painstaking attention to detail. In the process, they have more than earned the right to the title of assistant editors of *Readings in Formal Epistemology*.

Copyright Acknowledgments

We would like to thank authors, editors, publishers, copyright holders, permissions officers, and other parties who allowed us to reprint the papers found in this volume.

Helzner, J. and Hendricks, V.F. (2010). "Agency and Interaction: What we are and what we do in formal epistemology," *Journal for the Indian Council of Philosophical Research*, vol. XXVII:2, 2010: 44–71, special issue on *Logic and Philosophy Today*, guest edited by Amitabha Gupta and Johan van Benthem.

Part I. Bayesian Epistemology

Ramsey, F.P. (1926) "Truth and Probability," in Ramsey, F.P. (1931), *The Foundations of Mathematics and other Logical Essays*, Ch. VII, p.156–198, edited by R.B. Braithwaite, London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company.

Jeffrey, R.C. (1968), "Probable Knowledge," in *The Problem of Inductive Logic*, ed. I. Lakatos, 166–180, Amsterdam: North-Holland. Courtesy of Edith Jeffrey.

Van Fraassen, B.C. (1995) "Fine-Grained Opinion, Probability, and the Logic of Full Belief." *Journal of Philosophical Logic* 24 (4).

Gaifman, H. (1986) "A theory of higher order probabilities," in *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 conference on Theoretical aspects of reasoning about knowledge,* pp. 275—292, Morgan Kaufmann Publishers Inc. (Monterey, California).

Levi, I. (1974), "*On Indeterminate Probabilities*," *The Journal of Philosophy,* 71, 391–418.

Glymour, C. (1981) "Why I'm not a Bayesian," excerpt from Glymour, C. (1981) *Theory and Evidence,* Chicago University Press, 63–93.

Skyrms, B. (1993) "A Mistake in Dynamic Coherence Arguments – Discussion." *Philosophy of Science*, 60(2):320–328.

Arntzenius, F. (2003), "Some problems for conditionalization and reflection," *The Journal of Philosophy*, Vol. C, No. 7, 356–371.

M. J. Schervish, T. Seidenfeld and J. B. Kadane (2004) "Stopping to Reflect," *The Journal of Philosophy*, Vol. 101, No. 6, 315–322.

Part II. Belief Change

Alchourron, C.E., Gardenfors, P., and Makinson, D. (1985) "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions." *Journal of Symbolic Logic,* 50(2)*:* 510–530. Reprinted with the permission of the copyright holders, the Association of Symbolic Logic.

Hansson, S.O. (1993) "Theory Contraction and Base Contraction Unified." *Journal of Symbolic Logic,* 58(2): 602–625. Reprinted with the permission of the copyright holders, the Association of Symbolic Logic.

Levi, I. How Infallible but Corrigible Full Belief is Possible, hitherto unpublished.

Rott, H. (1993) "Belief Contraction in the Context of the General Theory of Rational Choice." *Journal of Symbolic Logic*, 58(4): 1426–1450. Reprinted with the permission of the copyright holders, the Association of Symbolic Logic.

Spohn, W. (2009) "A Survey of Ranking Theory." In Franz Huber and Christoph Schmidt-Petri (eds.). *Degrees of Belief*. Dordrecht: Springer.

Part III. Decision Theory

Savage, L. (1972) "Allais's Paradox" *The Foundations of Statistics*, Dover Publications, Inc., New York, 101–103.

Seidenfeld, T. (1988), "Decision Theory without 'Independence' or without 'Ordering': What is the Difference," *Economics and Philosophy*, 4: 267–290.

Gilboa, I. and M. Marinacci (forthcoming) "Ambiguity and the Bayesian Paradigm," *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress of the Econometric Society*.

Schervish, M.J., Seidenfeld, T. and Kadane, J.B. (1990) "State-Dependent Utilities," *Journal of the American Statistical Association*, Vol. 85, No. 411, 840–847.

Gibbard, A. and Joyce, J.M. (1998) "Causal Decision Theory." In Salvador Barberà, Peter J. Hammond, and Christian Seidl, eds., *Handbook of Utility Theory*, Vol. 1: Principles, pp. 701–740. Dordrecht & Boston: Kluwer.

Tversky, A. and Kahneman, D. (1992). "Advances in prospect theory: Cumulative representation of uncertainty." *Journal of Risk and Uncertainty* 5: 297–323.

Part IV. Logics of Knowledge and Belief

Hintikka, J. (2007) "Epistemology without Knowledge and without Belief" in *Socratic Epistemology: Explorations of Knowledge-Seeking by Questioning*, Cambridge University Press.

Dretske, F.I. (1970) "Epistemic Operators." *The Journal of Philosophy*, Vol. 67, No. 24, 1007–1023.

Lewis, D. (1996) "Elusive Knowledge," *Australasian Journal of Philosophy*, Vol. 74(4), 549–567. Courtesy of Stephanie Lewis.

Nozick, R. (1981) "Knowledge and Skepticism" In *Philosophical Explanations,* Harvard University Press, 167–169, 172–179, 197–211, 679–690. Reprinted by permission of the publisher from "Knowledge and Skepticism," in PHILOSOPHICAL EXPLANATIONS by Robert Nozick, pp. 167–169, 172–179, 197–211, 679–690, Cambridge, Mass.: The Belknap Press of Harvard University Press, Copyright © 1981 by Robert Nozick.

Stalnaker, R. (2006) "On Logics of Knowledge and Belief," *Philosophical Studies* 120, 169–199.

Parikh, R. (2008) "Sentences, Belief and Logical Omniscience, or What Does Deduction Tell Us?." *The Review of Symbolic Logic*, 1(4). Reprinted with the permission of the copyright holders, the Association of Symbolic Logic.

Artemov, S.N. (2008) "The logic of justification." *The Review of Symbolic Logic*, 1(4):477–513. Reprinted with the permission of the copyright holders, the Association of Symbolic Logic.

Kelly, K. (2004) "Learning Theory and Epistemology" in *Handbook of Epistemology*, I. Niiniluoto, M. Sintonen, and J. Smolenski (eds.), Dordrecht: Kluwer.

Williamson, T. (2004) "Some Computational Constraints in Epistemic Logic," in *Logic, Epistemology and the Unity of Science*, S. Rahman et al (eds). Dordrecht: Kluwer Academic Publishers: 437–456.

Part V. Interactive Epistemology

Lewis, D. (1969) *Convention: A Philosophical Study*, Harvard University Press, 24–42 (excerpt). Courtesy of Stephanie Lewis.

Barwise, J. (1988) "Three Views of Common Knowledge." In *Proc. TARK'88*: 365–379, Morgan Kaufmann Publishers.

Baltag, A. and Smets, S. (2008) "A Qualitative Theory of Dynamic Interactive Belief Revision," in G. Bonanno, W. van der Hoek, M. Wooldridge (eds.), *Logic and the Foundations of Game and Decision Theory*, Texts in Logic and Games, Vol 3, 9–58, Amsterdam University Press.

Aumann, R. (1976) "Agreeing to Disagree," *Annals of Statistics* 4, 1236–1239.

Aumann, R. and Brandenburger, A. (1995) "Epistemic Conditions for Nash Equilibrium," *Econometrica,* Vol. 63, No. 5,1161–1180.

Stalnaker, R. (1996), "*Knowledge, belief and counterfactual reasoning in games*," *Economics and Philosophy* 12: 133–163

Halpern, J.Y., (2001) "*Substantive Rationality and Backward Induction*," *Games and Economic Behavior*, Elsevier, vol. 37(2), 425–435.

Contents

Contributors

Carlos E. Alchourrón (deceased) Universidad de Buenos Aires, Buenos Aires, Argentina

Horacio Arló-Costa (deceased) Carnegie Mellon University, Pittsburgh, PA, USA

Frank Arntzenius University of Oxford, Oxford, UK

Sergei Artemov Graduate Center CUNY, New York, NY, USA

Robert J. Aumann Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem, Israel

Alexandru Baltag ILLC, University of Amsterdam, The Netherlands

Jon Barwise (deceased) Stanford University, Stanford, CA, USA

Henrik Boensvang University of Copenhagen, Copenhagen, Denmark

Adam Brandenburger Stern School of Business, Tandon School of Engineering, NYU Shanghai, New York University, New York, NY, USA

Fred I. Dretske (deceased) University of Wisconsin, Madison, WI, USA

Haim Gaifman Columbia University, New York, NY, USA

Peter Gärdenfors Lund University, Lund, Sweden

Allan Gibbard University of Michigan, Ann Arbor, MI, USA

Itzhak Gilboa HEC, Paris, France

Tel-Aviv University, Tel Aviv, Israel

Clark Glymour Carnegie Mellon University, Pittsburgh, PA, USA

Joseph Y. Halpern Computer Science Department, Cornell University, Ithaca, NY, USA

Sven Ove Hansson Division of Philosophy, KTH, Stockholm, Sweden

Jeffrey Helzner AIG, New York, NY, USA

Vincent F. Hendricks Center for Information and Bubble Studies, University of Copenhagen, Copenhagen, Denmark

Jaakko Hintikka (deceased) Boston University, Helsinki, Finland

Richard C. Jeffrey (deceased) Princeton University, Boston, MA, USA

James M. Joyce University of Michigan, Ann Arbor, MI, USA

Joseph B. Kadane Departments of Statistics and Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

Daniel Kahneman Princeton University, Princeton, NJ, USA

Kevin T. Kelly Carnegie Mellon University, Pittsburgh, PA, USA

Isaac Levi Columbia University, New York, NY, USA

David Lewis (deceased) Princeton University, Princeton, NJ, USA

David Makinson London School of Economics, London, UK

Massimo Marinacci Università Bocconi, Milano, Italy

Lawrence S. Moss Mathematics Department, Indiana University, Bloomington, IN, USA

Robert Nozick (deceased) Harvard University, Boston, MA, USA

Rohit Parikh City University of New York, New York, NY, USA

Frank P. Ramsey (deceased) University of Cambridge, Cambridge, UK

Rasmus K. Rendsvig Lund University, Lund, Sweden

Center for Information and Bubble Studies, University of Copenhagen, Copenhagen, Denmark

Hans Rott Department of Philosophy, University of Regensburg, Regensburg, Germany

Leonard Savage (deceased) Princeton University, New York, NY, USA

Mark J. Schervish Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA

Teddy Seidenfeld Departments of Philosophy and Statistics, Carnegie Mellon University, Pittsburgh, PA, USA

Brian Skyrms Department of Philosophy, University of California, Irvine, CA, USA

Sonja Smets ILLC, University of Amsterdam, The Netherlands

Sławomir Solecki Mathematics Department, Indiana University, Bloomington, IN, USA

Wolfgang Spohn Fachbereich Philosophie, Universität Konstanz, Konstanz, Germany

Robert Stalnaker Department of Linguistics and Philosophy, MIT, Cambridge, MA, USA

Johan van Benthem University of Amsterdam, Amsterdam, The Netherlands

Stanford University, Stanford, United States

Bas C. van Fraassen San Francisco State University, San Francisco, CA, USA

Amos Tversky (deceased) Stanford University, Stanford, CA, USA

Timothy Williamson University of Oxford, Oxford, UK

About the Editors

The late Horacio Arló-Costa was Professor of Philosophy at Carnegie Mellon University, Pennsylvania. Arló-Costa served as editor for the *Review of Symbolic Logic*, as area editor in epistemology for *Synthese*, and as member of the editorial board for the *Journal of Philosophical Logic*.

Vincent F. Hendricks is Professor of Formal Philosophy at the University of Copenhagen and Director of the Center for Information and Bubble Studies (CIBS). His recent publications include *Handbook of Formal Philosophy (2015), Infostorms* (2014), *Mainstream and Formal Epistemology* (2007), and *The Convergence of Scientific Knowledge* (2001). He served as editor-in-chief of *Synthese* from 2005 to 2015.

Johan van Benthem is University Professor of Logic at Amsterdam University, Henry Waldgrave Stuart Professor of Philosophy at Stanford University, and Distinguished Foreign Expert at Tsinghua University, Beijing. His recent publications include *Logical Dynamics of Information and Interaction* (2011), *Modal Logic for Open Minds* (2010), *Exploring Logical Dynamics* (1996), and *Language in Action* (1995). Van Benthem is coeditor, with Alice ter Meulen, of the *Handbook of Logic and Language* (1997).

Chapter 1
Agency and Interaction What We Are and What We Do in Formal Epistemology

Jeffrey Helzner and Vincent F. Hendricks

Introduction

Formal epistemology is a recent field of study in formal philosophy dating back only a couple of decades or so (Helzner and Hendricks 2011; Hendricks 2006). The point of departure of this essay is rooted in two philosophically fundamental and interrelated notions central to formal epistemology;

- *agency* – what agents are, and
- *interaction* – what agents do.

Agents may be individuals, or they may be groups of individuals working together. In each of the sections that follow, assumptions are made concerning the relevant features of the agents at issue. For example, such relevant features may include the agent's beliefs about its environment, its desires concerning various possibilities, the methods it employs in learning about its environment, and the strategies it adopts in its interactions with other agents in its environment. Fixing these features serves to bound investigations concerning interactions between the agent and its environment. The agent's beliefs and desires are assumed to inform its decisions. Methods employed by the agent for the purposes of learning are assumed to track or approximate or converge upon the facts of the agent's environment. Strategies adopted by the agent are assumed to be effective in some sense.

J. Helzner (✉)
AIG, New York, NY, USA
e-mail: jeffreyhelzner@yahoo.com

V.F. Hendricks
Center for Information and Bubble Studies, University of Copenhagen, Copenhagen, Denmark
e-mail: vincent@hum.ku.dk

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_1

In what follows is an attempt to locate predominant paradigms in formal epistemology – e.g., probability theory, belief revision theory, decision theory, logics of knowledge and belief and finally interactive epistemology – within the framework of agency and interaction.

Probability

Probabilities are very useful in formal epistemology. They are used to measure key epistemic components like belief and degrees thereof, the strength of expectations and predictions, may be used to describe actual occurrent frequencies in nature or in the agent's environment and of course probabilities play a paramount in accounting for notions of (Bayesian) confirmation (Earman 1992). Current cognitive models apply probabilities to represent aggregated experience in tasks involving language acquisition, problem solving and inductive learning, conditionalization and updating beliefs and scientific hypotheses.

What sorts of internal states are essential to the agent's representation of its environment? Various doxastic notions e.g., according to which the agent simply believes or is certain of propositions, in contrast to believing the proposition to some degree, are a traditional interest within mainstream epistemology. Some philosophers, e.g. Jeffrey (1992), have argued in favor of a doctrine known as *radical probabilism*. A central tenet of this doctrine is that various propositional attitudes of epistemic interest, especially full belief, are reducible to credal judgments. There are several ways that one might attempt such a reduction. Perhaps the most obvious is to identify full belief with maximal partial belief. For example, if it is assumed that the agent's credal state can be represented by a probability measure, then such a reduction would identify those propositions that are fully believed by the agent with those propositions that have maximal probability according to this representing measure. Following this proposal, it would seem that a proposition counts as a serious possibility for the agent just in case its negation is not assigned maximal probability according to the probability measure representing the agent's credal judgments. Hence, by the probability axioms, a proposition counts as seriously possible for the agent just in case it has nonzero probability under the representing measure. This leads to certain difficulties. For example, if the agent is concerned to estimate the height of an object that is sufficiently distant, then the agent might regard a continuum of values as possible – e.g., the height of the object is judged to be between three and four feet. According to the suggested reduction, such a continuum of possible values for the height of the object could not serve as a set of serious possibilities, since it is a mathematical fact that no probability measure can distribute positive probability to each element of such a continuum. The interested reader is urged to consult van Fraassen (1995) and Arlo-Costa (2001) for more sophisticated versions of probabilism.

Following Levi (1980), one may assume that the agent is, at each point in time, committed to full belief in some set of propositions concerning its environment.

Where the agent is not committed to full belief in a given proposition, the negation of that proposition is a serious possibility for the agent. The agent may judge some serious possibilities to be more probable than others. What can be said about these judgments? The received view, following a tradition that goes back to the work of Ramsey (1931), maintains that such credal judgments ought to be representable by a probability measure. This view has been criticized as being too weak and as being too strong. As for being too weak, the simple requirement that such judgments be representable by a probability measure says little about the extent to which these subjective probabilities should approximate objective probabilities, e.g., limiting frequencies in the sense of von Mises (1957) or perhaps even propensities in the sense of Popper (1959). Various principles have been offered in order to require that the subjective probabilities of a rational agent are informed by that agent's knowledge of objective probabilities (Kyburg 1974; Levi 1978; Lewis 1980). As for being too strong, requiring credal judgments to be representable by a probability measure implies, among other things, that such credal judgments are complete. A consequence of such a requirement is that, for any given pair of serious possibilities, the agent either judges one of the possibilities to be more probable than the other or the agent regards the possibilities as being equally probable. Thus, the requirement bars situations in which the agent, because of a lack of information, is unable to supply such a judgment. Such considerations, which to some extent echo earlier, related concerns of Keynes (1921) and Knight (1921), have motivated some – e.g., Kyburg (1968), Levi (1974) and Walley (1990) – to consider indeterminate probabilities, either in the form of interval-valued measures or sets of traditional measures, in representing rational credences.

Belief Change

As already hinted, some probability theorists tend to think that belief, as opposed to knowledge, may be good enough for action, deliberation and decision. Thus beliefs may suffice as they can serve important epistemic purposes while holding the information, expectations and conjectures that agents act on. Beliefs may be used for making creative leaps beyond what is logically implied by available information, evidence or knowledge and are crucial in agent interaction models representing what agents think about moves, strategies, payoffs and beliefs of other agents and what agent rationality amounts to Finally, beliefs and belief revision are prime vehicles for understanding the mechanism of learning by trial-and-error, one of the main motors of scientific inquiry in general.

Initially, an agent has beliefs about the environment with which it interacts. Sometimes these interactions are such that the agent, on pain of irrationality, must revise its beliefs. The classic example is that of a scientific agent who has beliefs about the world that might need to be revised in light of new data. The study of this sort of example has a long history in the philosophy of science, where it is often discussed at a relatively informal level in connection with topics such

as underdetermination. In the context of formal epistemology, the study of belief revision has been generalized to include various sorts of epistemic agents. Questions such as the following suggest the range of theoretical options that are available in connection with such investigations:

How are the potential belief states to be interpreted? One might take the belief states to represent partial beliefs; e.g., the agent has a certain degree of belief in proposition *P*. Alternatively, one might be interested in states of full belief, expectation or plain belief; e.g., the agent fully believes *P*, expects *P*, etc. Further refinements have been considered. For example, one might consider those full beliefs with respect to which the agent manifests some level of awareness; e.g., I am aware of my belief that I am presently writing the words of this sentence. In contrast to a focus on conscious beliefs, one might consider those propositions that the agent is committed to fully believing; e.g., all of those propositions that are deducible from my conscious beliefs.

How are the potential belief states to be represented? The answers to this question depend, at least to some extent, on how the previous question is answered. For example, if partial beliefs are the issue, then probability distributions might be taken as the basis for the representation so that a potential belief state is represented as a probability measure over the possible states of nature. On the other hand, if the problem is the representation of commitments to full belief (expectation, plain belief), then one might specify a suitably formalized language and represent each potential belief state as a theory formulated over the given language so that membership in the theory indicates full belief.

How are revisions interpreted? If credal states are the concern, then modifications of the credal state might be understood in terms of something like conditionalization. The interested reader is urged to consult (Halpern 2003) for a survey of various proposals concerning the representation and modification of credal states. What about revising or contracting states of full belief? When an instance of belief contraction concerning full beliefs is the result of the agent selecting from a set of (full) belief states that the agent recognizes as potential alternatives, then such an instance may be regarded as the resolution of a decision problem. Isaac Levi has developed a decision-theoretic approach to belief change; important discussions of Levi's approach include (1980), which considers belief change in the context Levi's general approach to epistemology, and Arlo-Costa and Levi (2006), Arlo-Costa and Liu (2011) which gives greater emphasis to the formal details concerning Levi's approach. Different connections between choice and belief revision are emphasized in Rott (1993). Rott demonstrates an important correspondence between the "AGM" account of belief revision offered in Alchourron et al. (1985) and the economists' study of rational choice functions. Finally, it is worth noting that where both partial and full beliefs are considered, there may be significant dependencies between the modification of these two sorts of belief states. For example, if the credal judgments of rational agents are a function of their judgments of full belief, as some philosophers assume, then changes to the latter may result in changes to the former.

There are other alternative interpretations of doxastic change. Spohn have considered change from the point of view of entire epistemic states rather than mere

beliefs in terms of ranking functions and plausibility representations (Spohn 2009) while Hansson have considered change from the point of belief bases as finite sets of sentences that likewise are possible axiomatic bases for a given theory (Hansson 1993).

Decision Theory

An agent interacts with its environment through the choices it makes. Choice presupposes alternatives, and a theory of rational choice should, at least, distinguish some of the available alternatives as admissible. As an example, consider those accounts of rational choice that are built on the concept of preference. One such account assumes that the agent has complete and transitive preferences over the set of available alternatives. Those alternatives that are optimal with respect to the given preference ranking are taken as admissible. This abstract preference-based account says nothing about the way in which preferences are informed by the agent's beliefs about its environment. Subjective expected utility theory [SEU], which is at the center of modern-day decision theory, provides significantly more detail than the abstract theory of preference optimization. SEU assumes that alternatives are acts, which, following Savage's classic formulation of SEU in Savage (1972), are functions from states to consequences. Drawing upon the earlier work of Ramsey (1931) on subjective probability and the work of von Neumann and Morgenstern (1947) on utility, Savage provides conditions on the agent's preferences over acts that guarantee the existence of a probability measure p and a utility function u such that the agent's preferences can be regarded as if they were the result of maximizing utility u with respect to probability p. According to the intended interpretation, the probability measure p represents the agent's degrees of belief concerning the possible states and the utility function u represents the extent to which the agent values the possible consequences.

The assumptions of SEU may be questioned in various ways. We focus on two ways that have generated significant interest among philosophers. First, why should it be that the rational agent's degrees of belief can be represented by a probability distribution p? As already noted, it is not clear why such an assumption should obtain in cases where the agent has very little information concerning the possible states. Second, in SEU it is assumed that the agent's subjective probability concerning the states is independent of the act that is chosen. Some question this assumption and offer examples in which a modification of SEU that provides for such dependencies, through the use of conditional probabilities, is supposed to give an irrational recommendation. The first line of questioning has led some – e.g., Ellsberg (1961), Levi (1974, 1977), and Gardenfors and Sahlin (1982) – to use indeterminate probabilities in their normative accounts of decision making under uncertainty. The second line of questioning has led some – e.g., Gibbard and Harper (1978), Lewis (1981), and Joyce (1999) – to investigate causal decision theory.

Logics of Knowledge and Belief

What is now known as epistemic logic started with the study of proper axiomatizations for knowledge, belief, certainty and other epistemic attitudes. Hintikka inaugurated the field with his seminal book Hintikka (1962) which focuses on axiomatizing knowledge and belief in mainly mono-agent systems. Agents are syntactically represented as indices on epistemic operators in a formal logical language. From the semantic perspective, to be an agent is to be an index on an accessibility relation between possible worlds representing the epistemic alternatives over which the agent has to succeed in order to know some proposition (interesting alternative semantics to Kripke semantics have been developed by Arlo-Costa and Pacuit 2006, Baltag and Moss 2004 and others). Like many other philosophical logics in their infancy, interesting axiomatizations governing the logics of knowledge and belief took center stage in the beginning together with nailing down important logical properties for these new logics. The field was living a somewhat isolated life remote from the general concerns of mainstream epistemology. Hintikka himself (and a few others like Lenzen 1978) was a notable exception and insisted on telling a better story, not about what agents are in the logical language, but about what they do and the meaning of epistemic axioms for epistemology (Stalnaker 2006). Accordingly, Hintikka took axioms of epistemic logic to describe a certain kind of strong rationality much in sync with the autoepistemological tradition of G.E. Moore and especially Norman Malcolm. Axioms of epistemic logic are really prescriptions of rationality in mono-agent systems. Epistemic logic has since been used address a number of important philosophical problems including for instance the Fitch Paradox (Brogaard and Salerno 2009), the problem of logical omniscience (Duc 1997; Parikh 2005), and various conceptual characterizations of knowledge and other epistemic attitudes (Kraus and Lehmann 1988).[1]

But rationality considerations are not only central to the singular agent acting in some environment, call it nature, but likewise, and perhaps especially, central to agents when in presence of other agents and interacting with these. Thus mono-agent systems had to be extended to multi-modal systems in order to get both agency and interaction off the epistemological ground for real. A sea-change took place in epistemic logic in the late 1980s and the beginning of the 1990s especially due to the work of Joseph Halpern and his collaborators (Fagin et al. 1995) and others (Meyer and Hoek 1995). Multiple agents were introduced into the logical language which, along with multiple epistemic accessibility relations on the semantic level, gave rise to a precise and adequate representation of the flow of information through an agent system, together with the nature of various protocols governing such systems. In this setting, possible worlds are to be understood as the states of the system taken as a whole, or sometimes the possible histories or consecutive runs of the system as a

[1] For solid overviews refer to De Bruin (2008) and Gochet and Gribomont (2006).

whole, that are compatible with the state transition directives which rule the system. Stalnaker has recently summarized the consequences of this sea-change precisely:

> The general lesson I drew from this work was that it was useful for epistemology to think of communities of knowers, exchanging information and interacting with the world, as (analogous to) distributed computer systems. (Hendricks and Roy 2010: 78)

Agent systems can now be thought of as encompassing everything from a group of robots on an assembly line to a group of poker players in Texas Hold 'Em. In turn, there is much more to what agents are nowadays, but also much more to what they do dynamically (as opposed to statically in terms of, say, (van Ditmarsch et al. 2008) epistemic axioms describing the rationality of single agents). Dynamic epistemic logic is a rich blend of studies ranging multi-agent axiomatizations of knowledge, belief, common knowledge and belief (Barwise 1988) certainty, uncertainty, doubt, ignorance and a host of other epistemic attitudes; models of the interplay between knowledge and games (Benthem 2001, 2007), knowledge and justification in mainstream epistemology (Artemov and Nogina 2005), social software (Parikh 2002), knowledge and public announcement of information (Baltag et al. 2002), knowledge intertwined with preferences, actions and decisions (Liu 2011); knowledge acquisition in light of formal learning theory, logical reliability, methods of scientific inquiry and computability studies (Gierasimczuk 2009; Hendricks 2001; Kelly 1996, 2004), belief revision (Baltag and Smets 2008), models of agent interaction in multi-agent systems; combined multi-agent and multi-modal systems in which for instance the development of knowledge over time may be scrutinized (Kraus and Lehmann 1988), relations between knowledge and deontic commitments investigated, divisions of cognitive labor modeled and so forth (for epistemic logic paired up with mainstream epistemological concerns, refer to Williamson (2006), Hendricks (2006) and Hendricks and Pritchard (2007)).

Interactive Epistemology

Theoretical economics is to a significant extent about understanding, anticipating and modeling phenomena like trading, stock speculation, real-estate dealing, hostile company take-overs, shareholding, convention and so forth. Obviously, agency and interaction play a paramount role here and seen from this perspective economics is about multi-individual and collective action balancing information and utility.

Independently, but informed by the developments in epistemic logic, economists have used game theory to scrutinize an extensive spread of the mentioned phenomena. By way of example, in 1976 the later Nobel Prize Laureate Robert Aumann published his famous Agreement Theorem in "Agreeing to Disagree" in which he describes conditions under which two "like minded" agents or players cannot "agree to disagree" in the sense that if the two players' posteriors of some event are common knowledge then they must coincide. In other words, in order to make trade

possible, agents have to agree to disagree (Aumann 1976). That is agency in terms of players, interaction in terms of games.

On the way to this result Aumann made a host of assumptions about the nature knowledge much in tune with what is to be found in epistemic logic like the axiomatic strength of knowledge in order to infer the backwards induction equilibrium and assumptions about what is common knowledge among the players (Halpern 2001). In 1999, Aumann coined a term for these kinds of study in theoretical economics: "Interactive epistemology" (Aumann 1999). It denotes an epistemic program studying shared knowledge and belief given more than one agent or player in an environment and has, as already suggested, strong ties to game theoretic reasoning and questions of common knowledge and belief, backward induction, various forms of game equilibria and strategies in games, (im)perfect information games, (bounded) rationality etc (Aumann and Brandenburger 1995; Stalnaker 1996, 2006).

Given its inauguration with Aumann, the program was in the beginning dominated by scholars drawn from theoretical economics and computer science rather than philosophy and logic, but recently philosophers and logicians have begun to pay close attention to what is going on in this striving program of formal epistemology. And for good reason too; social epistemology focuses on knowledge acquisition and justification in groups or institutions (Goldman 1999) and the extent to which exactly institutions may be viewed as genuine agents (List and Pettit 2011) while the interactive epistemological approach to agency and interaction also have close shaves with the major new focal points in dynamic epistemic logic (Benthem 2011) and much of the technical machinery is a common toolbox for both paradigms (Brandenburger 2007).

Formal Epistemology

Formal epistemology is the study of crucial concepts in mainstream epistemology including knowledge, belief (-change), certainty, rationality, reasoning, decision, justification, learning, agent interaction and information processing using formal tools from three streams; probability, logic and computability. In particular, the tools may come from tool boxes like modal logic, probability calculus, game theory, decision theory, formal learning theory, computability theory and distributed computing. Practitioners of formal epistemology include philosophers, computer scientists, social scientists, cognitive psychologists, theoretical economists, mathematicians, and theoretical linguists but also scholars from the empirical sciences like cognitive science, engineering and biology are onboard. This mixed bag of practitioners is surely a witness to the thoroughly interdisciplinary nature of formal epistemology and its wide range of applications in natural science, social science, humanities and the technical sciences.

Formal epistemology is right in the middle; between mainstream epistemology's fairly abstract theories on the one hand and the more concrete cognitive sciences

devoted to the empirical reality of agency and interaction on the other. In formal epistemology we are walking the fine line between theory and reality. This is as it should be: The hallmark of a progressive research program.

This is an edited and reorganized version of the paper "Agency and Interaction: What We Are and What We Do in Formal Epistemology", *Journal for the Indian Council of Philosophical Research*, nr. 2, vol. XXVII, 2010: 44–71, special issue on *Logic and Philosophy Today*, guest edited by Amitabha Gupta and Johan van Benthem.

References

Alchourron, C., Gardenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic, 50*, 510–530.

Arlo-Costa, H. (2001). Bayesian epistemology and epistemic conditionals: On the status of the export-import laws. *The Journal of Philosophy, 98*(11), 555–593.

Arlo-Costa, H., & Levi, I. (2006). Contraction: On the decision-theoretical origins of minimal change and entrenchment. *Synthese, 152*(1), 129–154.

Arlo-Costa, H., & Liu, H. (2011). Value-based contraction: A representation result. In *Proceedings of TARK'11*. New York: ACM Digital Library (ACM DL).

Arlo-Costa, H., & Pacuit, E. (2006). First order classical modal logic. *Studia Logica, 84*(2), 171–210.

Artemov, S., & Nogina, E. (2005). Introducing justification into epistemic logic. *Journal of Logic and Computation, 15*, 1059–1073.

Aumann, R. (1976). Agreeing to disagree. *The Annals of Statistics, 4*(6), 1236–1239.

Aumann, R. (1999). Interactive epistemology I. *International Journal of Game Theory, 28*(3), 263–300.

Aumann, R., & Brandenburger, A. (1995). Epistemic conditions for nash equilibrium. *Econometrica, 63*(5), 1161–1180.

Baltag, A., & Moss, L. (2004). Logics for epistemic programs. *Synthese/Knowledge, Rationality and Action, 139*, 165–224.

Baltag, A., & Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. In *Logic and the foundation of game and decision theory* (LOFT7 2008) (Vol. 3, pp. 13–60). Amsterdam: Amsterdam University Press.

Baltag, A., Moss, L. S., & Solecki, S. (2002). The logic of public annoucements, common knowledge, and private suspicion. In *Proceedings of TARK 1998* (pp. 43–56). Los Altos: Morgan Kaufmann Publishers.

Barwise, J. (1988). Three theories of common knowledge. In *Proceedings of the 2nd conference on theoretical aspects of reasoning about knowledge* (pp. 365–379). Pacific Grove: California.

Benthem, J. v. (2001). Games in dynamic epistemic logic. *Bulletin of Economic Research, 53*(4), 219–224.

Benthem, J. v. (2007). Logic games, from tools to models of interaction. In A. Gupta, R. Parikh, & J. van Benthem (Eds.), *Logic at the crossroads* (pp. 283–317). Mumbai: Allied Publishers.

Benthem, J. v. (2011). *Logical dynamics of information and interaction*. Cambridge: Cambridge University Press.

Brandenburger, A. (2007). The power of paradox: Some recent developments in interactive epistemology. *International Journal of Game Theory, 35*(4), 465–492.

Brogaard, B., & Salerno, J. (2009). Fitch's paradox of knowability. In E. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Palo Alto: Stanford University.

De Bruin, B. (2008). Epistemic logic and epistemology. In V. F. Hendricks & D. Prichard (Eds.), *New waves in epistemology*. London: Palgrave Macmillan.

Duc, H. N. (1997). Reasoning about rational, but not logically omniscient agents. *Journal of Logic and Computation, 7*, 633–648.
Earman, J. (1992). *Bayes or bust?* Cambridge: MIT.
Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics, 75*, 643–669.
Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning about knowledge*. Cambridge: MIT.
Gardenfors, P., & Sahlin, N. E. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese, 53*, 361–386.
Gibbard, A., & Harper, W. (1978). Counterfactuals and two kinds of expected utility". In C. A. Hooker, J. J. Leach, & E. F. McClennen (Eds.), *Foundations and applications of decision theory* (pp. 125–162). Dordrecht: Reidel.
Gierasimczuk, N. (2009). Learning by erasing in dynamic epistemic logic. In A. H. Dediu, A. M. Ionescu, & C. Martín-Vide *Language and automata theory and applications* (Lecture notes in computer science, Vol. 5457, pp. 362–373). Berlin/Heidelberg: Springer.
Gochet, P., & Gribomont, P. (2006). Epistemic logic. In *Logic and the modalities in the twentieth century* (Handbook of the history of logic, pp. 99–185). Amsterdam: Elsevier.
Goldman, A. (1999). *Knowledge in a social world*. New York: Oxford University Press.
Halpern, J. Y. (2001). Substantive rationality and backwards induction. *Games and Economic Behavior, 37*(2), 425–435.
Halpern, J. Y. (2003). *Reasoning about uncertainty.* Cambridge: MIT.
Hansson, S. O. (1993). Theory contraction and base contraction unified. *Journal of Symbolic Logic, 58*(2), 602–625.
Helzner, J., & Hendricks, V. F. (2011, forthcoming). Formal epistemology. In *Oxford bibliography online*. http://www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577-0140.xml?rskey=a6uyIx&result=81
Hendricks, V. F. (2001). *The convergence of scientific knowledge – A view from the limit*. Dordrecht: Springer.
Hendricks, V. F. (2006). *Mainstream and formal epistemology*. New York: Cambridge University Press.
Hendricks, V. F., & Pritchard, D. (Eds.) (2007). *Epistemology: 5 questions*. New York: Automatic Press/London: VIP.
Hendricks, V. F., & Roy, O. (Eds.) (2010). *Epistemic logic: 5 questions*. New York: Automatic Press/London: VIP.
Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Ithaca: Cornell University Press.
Jeffrey, R. (1992). *Probability and the art of judgment*. New York: Cambridge University Press.
Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.
Kelly, K. T. (1996). *The logic of reliable inquiry*. New York: Oxford University Press.
Kelly, K. T. (2004). Learning theory and epistemology. In I. Niiniluoto, M. Sintonen, & J. Smolenski (Eds.), *Handbook of epistemology*. Dordrecht: Kluwer.
Keynes, J. M. (1921). *A treatise on probability*. London: MacMillan.
Knight, F. H. (1921). *Risk, uncertainty and profit*. Boston/New York: Houghton-Mifflin.
Kraus, S., & Lehmann, D. (1988). Knowledge, belief and time. *Theoretical Computer Science, 58*, 155–174.
Kyburg, H. E. (1968). Bets and beliefs. *American Philosophical Quarterly, 5*(1), 54–63.
Kyburg, H. E. (1974). *The logical foundations of statistical inference*. Dordrecht: Reidel.
Lenzen, W. (1978). *Recent work in epistemic logic* (Acta philosophica fennica, Vol. 30, pp. 1–219). Amsterdam: North-Holland
Levi, I. (1974). On indeterminate probabilities. *The Journal of Philosophy, 71*, 391–418.
Levi, I. (1977). Direct inference. *The Journal of Philosophy, 74*, 5–29.
Levi, I. (1978). In A. Hooker, J. J. Leach, & E. F. McClennen (Eds.), *Foundations and applications of decision theory* (pp. 263–273). Dordrecht/Boston: D. Reidel.

Levi, I. (1980). *The enterprise of knowledge*. Cambridge: MIT.
Lewis, D. (1980). A subjectivist's guide to objective chance. In R. Jeffrey (Ed.), *Studies in inductive logic* (Vol. II). Berkeley/Los Angeles: University of California Press.
Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy, 59*, 5–30.
List, C., & Pettit, P. (2011). *Group agency*. New York: Oxford University Press.
Liu, F. (2011). *Reasoning about preference dynamics*. Dordrecht: Springer.
Meyer, J. -J. Ch., & Hoek, W. van der (1995). *Epistemic logic for AI and computer science* (Cambridge tracts in theoretical computer science, Vol. 41). Cambridge: Cambridge University Press.
Parikh, R. (2002). Social software. *Synthese, 132*, 187–211.
Parikh, R. (2005). Logical omniscience and common knowledge: WHAT do we know and what do WE know? In *Proceedings of TARK05* (pp. 62–77). Singapore: National University of Singapore.
Popper, K. R. (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science, 10*, 26–42.
Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.) *The foundations of mathematics and other logical essays*. London: Routledge and Kegan Paul.
Rott, H. (1993). Belief contraction in the context of the general theory of rational choice. *The Journal of Symbolic Logic, 58*(4), 1426–1450.
Savage, L. J. (1972) *The foundations of statistics*. New York: Dover. The Dover edition is a republication of the 1954 work.
Spohn, W. (2009). A survey of ranking theory. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (pp. 185–228). Dordrecht: Springer.
Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy, 12*, 133–163.
Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies, 128*(1), 169–199.
van Ditmarsch, H., Hoek, W. v. d., & Kooi, B. (2008). *Dynamic epistemic logic*. Dordrecht: Springer.
van Fraassen, B. (1995). Fine-grained opinion, probability, and the logic of full belief. *Journal of Philosophical Logic, 24*(4), 349–377.
von Mises, R. (1957). *Probability, statistics and truth* (Revised English ed.). New York: Macmillan.
von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton: Princeton University Press.
Walley, P. (1990). *Statistical reasoning with imprecise probabilities*. New York: Chapman and Hall.
Williamson, T. (2006). *Knowledge and its limits*. Oxford: Oxford University Press.

Part I
Bayesian Epistemology

Chapter 2
Introduction

Horacio Arló-Costa, Vincent F. Hendricks, and Johan van Benthem

There are various possible ways of articulating what Bayesian epistemology is and how it relates to other branches of formal and mainstream epistemology. Following the steps of Ramsey, Richard Jeffrey outlines in his article "Probable Knowledge" a possible way of constructing an epistemology grounded on Bayesian theory. While knowledge is a central notion in traditional epistemology (and in various branches of formal epistemology) Jeffrey suggests an epistemology where knowledge does not have the importance generally attributed to it. The idea is "[...] to try to make the concept of belief do the work that philosophers have generally assigned to the grander concept" (knowledge). Moreover the notion of belief is pragmatically analyzed along the lines proposed by Ramsey: "the kind of measurement of belief with which probability is concerned is .... a measurement of belief qua basis of action". The result of this move is to conceive the logic of partial belief as a branch of decision theory. So, the first two essays in this section are also quite relevant for the section of decision theory presented below (Ramsey's essay contains the first axiomatic presentation of decision theory). Both Jeffrey and Ramsey present the foundations of an epistemology which is deeply intertwined with a theory of action.

Horacio Arló-Costa was deceased at the time of publication.

H. Arló-Costa (deceased)
Carnegie Mellon University, Pittsburgh, PA, USA

V.F. Hendricks (✉)
Center for Information and Bubble Studies, University of Copenhagen, Copenhagen, Denmark
e-mail: vincent@hum.ku.dk

J. van Benthem
University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Stanford University, Stanford, United States
e-mail: johan@science.uva.nl

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_2

This move has a behaviorist pedigree but perhaps the behavioral inspiration is not an essential ingredient of an interpretation of the formal theory that thus arises.

The notion of certainty or full belief does not play a central role in Jeffrey's epistemology either. According to him we are only certain of mathematical and logical truths and the truths related immediately to experience. The rest is the domain of probable knowledge. To be coherent with this view Jeffrey has to propose a modification of the classical notion of learning by conditioning on data (which occupies a central role in various versions of Bayesian Epistemology as used in the social sciences like economics or psychology). In fact, according to conditioning when one learns a new piece of evidence this information acquires measure one in spite of being based on perfectly fallible evidence. A modification of conditioning that permits to update on uncertain evidence is presented in "Probable Knowledge". The version of Jeffrey's article reprinted here contains as well comments by L. Hurwicz and P. Suppes and responses by Jeffrey. Some of Suppes' comments point in the direction of constructing a theory of rationality that is sensible to our cognitive limitations. The possibility of constructing a "bounded" theory of rationality only started with the seminal work of Herb Simon (Rational choice and the structure of the environment, *Psychological Review*, Vol. 63 No. 2, 129–138. 1956) and is today an active area of investigation in economics, psychology and philosophy.

The uses of Bayesianism in epistemology are usually dismissed by realist philosophers for delivering a subjective picture of rationality that is not sufficiently sensible to the way in which the behavior of rational agents is connected with the structure of the environment. Simon's work was certainly sensible nevertheless to ecological considerations. And Ramsey's essay ends with programmatic ideas differentiating what he called "the logic of consistency" from the "logic of truth". Even Bruno de Finetti who is usually presented as a precursor of Jeffrey's radical probabilism, had philosophical ideas about certainty that clashed with this view (he thought that certainty has to be assumed as a primitive alongside probability, and that we can be certain of more than mere tautologies). Moreover his mature philosophical work veered towards a more objective point of view. For example he dismissed the use of Dutch Book arguments and embraced the use of foundational arguments in terms of scoring rules, a methodological move favored today by many "objective" Bayesians (a presentation of de Finetti's mature views appear in: *Philosophical Lectures on Probability*: collected, edited, and annotated by Alberto Mura, Synthese Library, Springer, 2008).

Van Fraassen introduces in his essay a version of radical probabilism (the term was coined by Jeffrey) where the only epistemological primitive is a notion of conditional probability. Van Fraassen sees this notion as encoding a notion of supposition from which he derives a non-trivial notion of full belief. According to this view it is perfectly possible to be sure of the contingent propositions of science and everyday knowledge. One can see van Fraassen's theory as introducing paradox-free acceptance rules that link probability and belief (some of the usual acceptance rules of this type like high probability rules are known to be the victim of various forms of paradoxes, like the paradox of the lottery first proposed by Henry Kyburg (*Probability and the Logic of Rational Belief*, Wesleyan University

Press, 1961)). Jeffrey renounced to the use of any form of acceptance rules of this type and therefore proposed to eliminate any notion of qualitative belief without a probabilistic origin. Van Fraassen has exactly the opposite intention: namely to tend bridges between traditional and Bayesian epistemology via the use of novel acceptance rules.

Most of the models of probability update considered above deal with synchronic or suppositional change. Is it possible to extend these models to cover cases of genuine changes of probabilistic belief? David Lewis, Bas van Fraassen and Paul Teller provided in the 1970s various dynamic coherence arguments showing that one should update diachronically via conditioning on pain of incoherence (see references in the section on "Further reading" below). If we denote by $P_t(B|A)$ the conditional probability of B given A at time t and by $P_t(.)$ the monadic probability P at time t, we can denote by $P_{t'}$ the monadic probability at time t' where the total evidence gathered between these two times is A. Then the idea that one should update diachronically via conditioning can be expressed formally by: $P_{t'}(B) = P_t(B|A)$. These arguments in favor of this diachronic principle are dynamic versions of the static Dutch Book arguments first proposed in Ramsey's essay. Unfortunately these arguments are considerably more controversial than the well know static Dutch Book argument. The article by Brian Skyrms summarizes almost 30 years of arguments pro and con dynamic book arguments and offers a temperate and more modest version of these arguments that he thinks is valid. This debate is nevertheless still open.

Levi's essay is embedded on his own version of Bayesian epistemology where the notion of full belief is taken as a primitive alongside probability. But the central goal of the article reprinted here is to study the case where probabilities are indeterminate, imprecise or vague. Levi also thinks that a theory of partial belief (precise or imprecise) should be conceived as a branch of decision theory and therefore proposes rules for deciding in conditions of uncertainty. Currently there is a fair amount of work in this area not only in philosophy but also in statistics, computer science and economics.

Gaifman's article focuses on characterizing the structure of higher order probability. In particular he investigates a form of the so-called Miller's principle by which a rational agent adjusts his or her probabilities in accordance to the probabilities of an expert. So, we have a principle of this form:

(Miller's Principle) $P_{you}(A \mid P_{expert}(A) = r) = r$[1]

van Fraassen proposed a diachronic version of this principle for a single agent:

(Reflection) $P_{now}(A \mid P_{later}(A) = r) = r$

[1] Actually Gaifman's formulation of the principle is formally cleaner. He defines $PR(A, r) = \{x \text{ in } W \mid p_x(A) = r\}$, where p_x is an expert function at world x, i.e. the distribution chosen by the expert at that world. Then he formulates the principle as follows: $P(A \mid PR(A, r)) = r$.

Arntzenius' article presents five puzzles showing that rational people can update their degrees of belief in manners that violate Bayesian conditioning and Reflection. But the article by M.J. Schervish, T. Seidenfeld and J. Kadane disputes that Arntzenius' examples impose any new restrictions or challenges to conditioning or Reflection beyond what is already familiar about these principles.

Suggested Further Reading

- An excellent introduction to Ramsey's philosophy in general and to the essay reprinted here in particular can be found in the corresponding chapters of: *The Philosophy of F.P. Ramsey*, by Nils-Eric Sahlin, Cambridge University Press, 2008. The classical introduction to Richard Jeffrey's decision theory is his: *The Logic of Decision*, University Of Chicago Press: 2nd edition (July 15, 1990). A detailed articulation of radical probabilism can be found in Probability and the Art of Judgment, Cambridge Studies in Probability, Induction and Decision Theory (Mar. 27, 1992). The theory of probability cores presented in van Fraassen's article has been slightly modified and extended in a paper by Horacio Arlo-Costa and Rohit Parikh: "Conditional Probability and Defeasible Inference," *Journal of Philosophical Logic* 34, 97-119, 2005. The best axiomatic presentation of primitive conditional probability is given by Lester E. Dubins in his article Finitely Additive Conditional Probabilities, Conglomerability and Disintegrations, *The Annals of Probability*, 3(1):89–99, 1975. Teddy Seidenfeld wrote an accessible note presenting recent results in this area in: Remarks on the theory of conditional probability: Some issues of finite versus countable additivity, *Probability Theory*, V.F. Hendricks et al. (eds.) 2001, pp. 167-178. Alan Hájek articulated a philosophical defense of the use of primitive conditional probability in: What Conditional Probability Could Not Be, *Synthese*, Vol. 137, No. 3, Dec., 2003. Finally there is an interesting article by David Makinson linking conditional probability and central issues in belief change: Conditional probability in the light of qualitative belief change, to appear in a 2011 issue of the *Journal of Philosophical Logic* marking 25 years of AGM. References to other classical articles in this area by Karl Popper, Alfred Renyi and Bruno de Finetti appear in the aforementioned articles.
- Brian Skyrms has also contributed to the theory of higher order probability. One accessible article is: "Higher Order Degrees of Belief," in D. H. Mellor (ed.), *Prospects for Pragmatism*. Cambridge: Cambridge University Press, 109–13. Isaac Levi has articulated his theory of indeterminate probabilities in various books and articles. One of the classical sources is: *The Enterprise of Knowledge*, MIT Press, Cambridge, 1983. More information about Levi's version of decision theory under uncertainty appears in section 7 on Decision Theory below.
- There are two classical sources for the formulation of dynamic Dutch books. One is: Teller, P. (1973), "Conditionalization and Observation", *Synthese* 26: 218-258. The other is: van Fraassen, Bas (1984), "Belief and the Will," *Journal of Philosophy* 81: 235–256. The second piece introduces also a theory of second order probability that complements the writings of Skyrms and Gaifman. Van Fraassen introduces there the Reflection Principle. The original formulation of some of the puzzles discussed by Arntzenius and Seidenfeld is a brief piece by Adam Elga: "Self-Locating Belief and the Sleeping Beauty problem, *Analysis*, 60(2): 143-147, 2000. More detailed reference to the work by Carnap on induction and confirmation can be found in the bibliography of Maher's paper. The so-called Raven's Paradox appeared for the first time in a seminal article by Carl Hempel: "Studies in the Logic of Confirmation (I.)," *Mind,* New Series, Vol. 54, No. 213 (Jan., 1945), pp. 1-26. Branden Fitelson and James Hawthorne offer an alternative and interesting Bayesian account of the paradox in: "How Bayesian Confirmation Theory Handles the Paradox of the Ravens," in E. Eels and J. Fetzer (eds.), *The Place of Probability in Science*, Chicago: Open Court. Further information about confirmation theory can be found in a classical book by John Earman: *Bayes or Bust? A Critical*

Examination of Bayesian Confirmation Theory, MIT Press, 1992. Another classical source is: *Scientific Reasoning: The Bayesian Approach*, by Colin Howson and Peter Urbach, Open Court; 3rd edition, 2005. A interesting book touching a cluster of issues recently discussed in this area like coherence and the use of Bayesian networks in epistemology is: *Bayesian Epistemology* by Luc Bovens and Stephan Hartmann, Oxford University Press,. 2004.

- Another important formal epistemological issue is investigated by Timothy Williamson in his paper, "Conditionalizing on Knowledge", *British Journal for the Philosophy of Science* 49 (1), 1998: 89-121, which intends to integrate the theory of probability and probability kinematics, with other epistemological notions like the notion of knowledge. The theory of *evidential probability* that thus arises is based on two central ideas: (1) the evidential probability of a proposition is its probability conditional on the total evidence (or conditional on evidence propositions); (2) one's total evidence is one's total knowledge. The tools of epistemic logic are used in order to represent the relevant notion of knowledge.
- Jeffrey does not adopt (1) but according to his modified notion of updating once a proposition has evidential probability 1, it keeps it thereafter (monotony). This is a feature shared by Jeffrey's updating and the classical notion of updating. Williamson does embrace (1) but develops a model of updating that abandons monotony. This seems a very promising strategy given the limited applicability of a cumulative model of growth of knowledge. Similarly motivated models (that are nevertheless formally quite different) have been proposed by Isaac Levi, Peter Gärdenfors. Gärdenfors' model appears in his book *Knowledge in Flux* (see the corresponding reference in the bibliographical references of chapter 6). Levi presents his account in *The Enterprise of Knowledge* (the reference appears in the bibliographical section below). Both models appeal directly not only to qualitative belief but also to models of belief change (*contraction* and *revision* - see chapter 6).
- Philosophers of science have traditionally appealed to Bayesian theory in order to provide a Carnapian explication of the notoriously vague, elusive and paradox-prone notion of *confirmation* or *partial justification* in science. Patrick Maher revives in his article, "Probability Captures the Logic of Scientific Confirmation," in *Contemporary Debates in the Philosophy of Science*, ed. Christopher Hitchcock, Blackwell, 69–93, the Carnapian program of inductive inference in order to provide one of these explications. In contrast Clark Glymour and Kevin Kelly argue in their article, "Why Probability Does Not Capture the Logic of Scientific Justification", in Christopher Hitchcock, ed., *Contemporary Debates in the Philosophy of Science*, London: Blackwell, 2004, that Bayesian confirmation cannot deliver the right kind of account of the logic of scientific confirmation. One of the reasons for this skepticism is that they think that scientific justification should reflect how intrinsically difficult is to find the truth and how efficient one's methods are at finding it. So, their skepticism arises because they think that Bayesian confirmation captures neither aspect of scientific justification. While deploying their arguments the two articles discuss the well-known paradox of confirmation first proposed by Hempel, Carnap's research program on the philosophy of probability and induction and the possible application of learning theory in order to offer a non-Bayesian account of scientific justification. The article by Glymour and Kelly continues Glymour's earlier critique of the applications of Bayesianism in philosophy of science (also reprinted here). This earlier piece contains the original versions of some influential and much-discussed conundra engendered by Bayesian confirmation (like the problem of Old Evidence).

Chapter 3
Truth and Probability

Frank P. Ramsey

> To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is and of what is not that it is not is true.
> – *Aristotle*.

> When several hypotheses are presented to our mind which we believe to be mutually exclusive and exhaustive, but about which we know nothing further, we distribute our belief equally among them .... This being admitted as an account of the way in which we actually do distribute our belief in simple cases, the whole of the subsequent theory follows as a deduction of the way in which we must distribute it in complex cases if we *would be consistent*.
> – *W. F. Donkits*.

> The object of reasoning is to find out, from the consideration of what we already know, something else which we do not know. Consequently, reasoning is good if it be such as to give a true conclusion from true premises, and not otherwise.
> – *C. S. Peirce*.

> Truth can never be told so as to be understood, and not be believed.
> – *W. Blake*.

Foreword

In this essay the Theory of Probability is taken as a branch of logic, the logic of partial belief and inconclusive argument; but there is no intention of implying that this is the only or even the most important aspect of the subject. Probability is of fundamental importance not only in logic but also in statistical and physical science, and we cannot be sure beforehand that the most useful interpretation of it

Frank P. Ramsey was deceased at the time of publication.

F.P. Ramsey (deceased)
University of Cambridge, Cambridge, UK

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_3

in logic will be appropriate in physics also. Indeed the general difference of opinion between statisticians who for the most part adopt the frequency theory of probability and logicians who mostly reject it renders it likely that the two schools are really discussing different things, and that the word 'probability' is used by logicians in one sense and by statisticians in another. The conclusions we shall come to as to the meaning of probability in logic must not, therefore, be taken as prejudging its meaning in physics.

The Frequency Theory

In the hope of avoiding some purely verbal controversies, I propose to begin by making some admissions in favour of the frequency theory. In the first place this theory must be conceded to have a firm basis in ordinary language, which often uses 'probability' practically as a synonym for proportion; for example, if we say that the probability of recovery from smallpox is three-quarters, we mean, I think, simply that that is the proportion of smallpox cases which recover. Secondly, if we start with what is called the calculus of probabilities, regarding it first as a branch of pure mathematics, and then looking round for some interpretation of the formulae which shall show that our axioms are consistent and our subject not entirely useless, then much the simplest and least controversial interpretation of the calculus is one in terms of frequencies. This is true not only of the ordinary mathematics of probability, but also of the symbolic calculus developed by Mr. Keynes; for if in his $\frac{a}{h}$, a and h are taken to be not propositions but propositional functions or class-concepts which define finite classes, and $\frac{a}{h}$ is taken to mean the proportion of members of h which are also members of a, then all his propositions become arithmetical truisms.

Besides these two inevitable admissions, there is a third and more important one, which I am prepared to make temporarily although it does not express my real opinion. It is this. Suppose we start with the mathematical calculus, and ask, not as before what interpretation of it is most convenient to the pure mathematicism, but what interpretation gives results of greatest value to science in general, then it may be that the answer is again an interpretation in terms of frequency; that probability as it is used in statistical theories, especially in statistical mechanics – the kind of probability whose logarithm is the entropy – is really a ratio between the numbers, of two classes, or the limit of such a ratio. I do not myself believe this, but I am willing for the present to concede to the frequency theory that probability as used in modern science is really the same as frequency.

But, supposing all this admitted, it still remains the case that we have the authority both of ordinary language and of many great thinkers for discussing under the heading of probability what appears to be quite a different subject, the logic of partial belief. It may be that, as some supporters of the frequency theory have maintained, the logic of partial belief will be found in the end to be merely the study of frequencies, either because partial belief is definable as, or by reference to, some

sort of frequency, or because it can only be the subject of logical treatment when it is grounded on experienced frequencies. Whether these contentions are valid can, however, only be decided as a result of our investigation into partial belief, so that I propose to ignore the frequency theory for the present and begin an inquiry into the logic of partial belief. In this, I think, it will be most convenient if, instead of straight away developing my own theory, I begin by examining the views of Mr Keynes, which are so well known and in essentials so widely accepted that readers probably feel that there is no ground for re-opening the subject *de novo* until they have been disposed of.

Mr. Keynes' Theory

Mr Keynes[1] starts from the supposition that we make probable inferences for which we claim objective validity; we proceed from full belief in one proposition to partial belief in another, and we claim that this procedure is objectively right, so that if another man in similar circumstances entertained a different degree of belief, he would be wrong in doing so. Mr Keynes accounts for this by supposing that between any two propositions, taken as premiss and conclusion, there holds one and only one relation of a certain sort called probability relations; and that if, in any given case, the relation is that of degree α, from full belief in the premiss, we should, if we were rational, proceed to a belief of degree α in the conclusion.

Before criticising this view, I may perhaps be allowed to point out an obvious and easily corrected defect in the statement of it. When it is said that the degree of the probability relation is the same as the degree of belief which it justifies, it seems to be presupposed that both probability relations, on the one hand, and degrees of belief on the other can be naturally expressed in terms of numbers, and then that the number expressing or measuring the probability relation is the same as that expressing the appropriate degree of belief. But if, as Mr. Keynes holds, these things are not always expressible by numbers, then we cannot give his statement that the degree of the one is the same as the degree of the other such a simple interpretation, but must suppose him to mean only that there is a one-one correspondence between probability relations and the degrees of belief which they justify. This correspondence must clearly preserve the relations of greater and less, and so make the manifold of probability relations and that of degrees of belief similar in Mr Russell's sense. I think it is a pity that Mr Keynes did not see this clearly, because the exactitude of this correspondence would have provided quite as worthy material scepticism as did the numerical measurement of probability relations. Indeed some of his arguments against their numerical measurement appear to apply quite equally well against their exact correspondence with degrees of belief; for instance, he argues that if rates of insurance correspond to subjective, i.e. actual, degrees of

[1] J.M. Keynes, *A Treatise on Probability* (1921).

belief, these are not rationally determined, and we cannot infer that probability relations can be similarly measured. It might be argued that the true conclusion in such a case was not that, as Mr Keynes thinks, to the non-numerical probability relation corresponds a non-numerical degree of rational belief, but that degrees of belief, which were always numerical, did not correspond one to one with the probability relations justifying them. For it is, I suppose, conceivable that degrees of belief could be measured by a psychogalvanometer or some such instrument, and Mr Keynes would hardly wish it to follow that probability relations could all be derivatively measured with the measures of the beliefs which they justify.

But let us now return to a more fundamental criticism of Mr Keynes' views, which is the obvious one that there really do not seem to be any such things as the probability relations he describes. He supposes that, at any rate in certain cases, they can be perceived; but speaking for myself I feel confident that this is not true. I do not perceive them, and if I am to be persuaded that they exist it must be by argument; moreover I shrewdly suspect that others do not perceive them either, because they are able to come to so very little agreement as to which of them relates any two given propositions.

All we appear to know about them are certain general propositions, the laws of addition and multiplication; it is as if everyone knew the laws of geometry but no one could tell whether any given object were round or square; and I find it hard to imagine how so large a body of general knowledge can be combined with so slender a stock of particular facts. It is true that about some particular cases there is agreement, but these somehow paradoxically are always immensely complicated; we all agree that the probability of a coin coming down heads is $\frac{1}{2}$, but we can none of us say exactly what is the evidence which forms the other term for the probability relation about which we are then judging. If, on the other hand, we take the simplest possible pairs of propositions such as 'This is red' and 'That is blue' or 'This is red' and 'That is red', whose logical relations should surely be easiest to see, no one, I think, pretends to be sure what is the probability relation which connects them. Or, perhaps, they may claim to see the relation but they will not be able to say anything about it with certainty, to state if it is more or less than $\frac{1}{3}$, or so on. They may, of course, say that it is incomparable with any numerical relation, but a relation about which so little can be truly said will be of little scientific use and it will be hard to convince a sceptic of its existence. Besides this view is really rather paradoxical; for any believer in induction must admit that between 'This is red' as conclusion and 'This is round', together with a billion propositions of the form 'a is round and red' as evidence, there is a finite probability relation; and it is hard to suppose that as we accumulate instances there is suddenly a point, say after 233 instances, at which the probability relation becomes finite and so comparable with some numerical relations.

It seems to me that if we take the two propositions 'a is red', 'b is red', we cannot really discern more than four simple logical relations between them; namely identity of form, identity of predicate, diversity of subject, and logical independence of import. If anyone were to ask me what probability one gave to the other, I should not try to answer by contemplating the propositions and trying to discern a logical

relation between them, I should, rather, try to imagine that one of them was all that I knew, and to guess what degree of confidence I should then have in the other. If I were able to do this, I might no doubt still not be content with it, but might say 'This is what I should think, but, of course, I am only a fool' and proceed to consider what a wise man would think and call that the degree of probability. This kind of self-criticism I shall discuss later when developing my own theory; all that I want to remark here is that no one estimating a degree of probability simply contemplates the two propositions supposed to be related by it; he always considers *inter alia* his own actual or hypothetical degree of belief. This remark seems to me to be borne out by observation of my own behaviour; and to be the only way of accounting for the fact that we can all give estimates of probability in cases taken from actual life, but are quite unable to do so in the logically simplest cases in which, were probability a logical relation, it would be easiest to discern.

Another argument against Mr Keynes' theory can, I think, be drawn from his inability to adhere to it consistently even in discussing first principles. There is a passage in his chapter on the measurement of probabilities which reads as follows: –

> Probability is, *vide* Chapter 11 (§12), relative in a sense to the principles of *human* reason. The degree of probability, which it is rational for *us* to entertain, does not presume perfect logical insight, and is relative in part to the secondary propositions which we in fact know; and it is not dependent upon whether more perfect logical insight is or is not conceivable. It is the degree of probability to which those logical processes lead, of which our minds are capable; or, in the language of Chapter II, which those secondary propositions justify, which we in fact know. If we do not take this view of probability, if we do not limit it in this way and make it, to this extent, relative to human powers, we are altogether adrift in the unknown; for we cannot ever know what degree of probability would be justified by the perception of logical relations which we are, and must always be, incapable of comprehending.[2]

This passage seems to me quite unreconcilable with the view which Mr Keynes adopts everywhere except in this and another similar passage. For he generally holds that the degree of belief which we are justified in placing in the conclusion of an argument is determined by what relation of probability unites that conclusion to our premisses, There is only one such relation and consequently only one relevant true secondary proposition, which, of course, we may or may not know, but which is necessarily independent of the human mind. If we do not know it, we do not know it and cannot tell how far we ought to believe the conclusion. But often, he supposes, we do know it; probability relations are not ones which we are incapable of comprehending. But on this view of the matter the passage quoted above has no meaning: the relations which justify probable beliefs are probability relations, and it is nonsense to speak of them being justified by logical relations which we are, and must always be, incapable of comprehending. The significance of the passage for our present purpose lies in the fact that it seems to presuppose a different view of probability, in which indefinable probability relations play no part, but in which the degree of rational belief depends on a variety of logical relations. For instance, there

[2] p. 32, his italics.

might be between the premiss and conclusion the relation that the premiss was the logical product of a thousand instances of a generalization of which the conclusion was one other instance, and this relation, which is not an indefinable probability relation but definable in terms of ordinary logic and so easily recognizable, might justify a certain degree of belief in the conclusion on the part of one who believed the premiss. We should thus have a variety of ordinary logical relations justifying the same or different degrees of belief. To say that the probability of a given h was such-and-such would mean that between a and h was some relation justifying such-and-such a degree of belief. And on this view it would be a real point that the relation in question must not be one which the human mind is incapable of comprehending.

This second view of probability as depending on logical relations but not itself a new logical relation seems to me more plausible than Mr Keynes' usual theory; but this does not mean that I feel at all inclined to agree with it. It requires the somewhat obscure idea of a logical relation justifying a degree of belief, which I should not like to accept as indefinable because it does not seem to be at all a clear or simple notion. Also it is hard to say what logical relations justify what degrees of belief, and why; any decision as to this would be arbitrary, and would lead to a logic of probability consisting of a host of so-called 'necessary' facts, like formal logic on Mr Chadwick's view of logical constants.[3] Whereas I think it far better to seek an explanation of this 'necessity' after the model of the work of Mr Wittgenstein, which enables us to see clearly in what precise sense and why logical propositions are necessary, and in a general way why the system of formal logic consists of the propositions it does consist of, and what is their common characteristic. Just as natural science tries to explain and account for the facts of nature, so philosophy should try, in a sense, to explain and account for the facts of logic; a task ignored by the philosophy which dismisses these facts as being unaccountably and in an indefinable sense 'necessary'.

Here I propose to conclude this criticism of Mr Keynes' theory, not because there are not other respects in which it seems open to objection, but because I hope that what I have already said is enough to show that it is not so completely satisfactory as to render futile any attempt to treat the subject from a rather different point of view.

Degrees of Belief

The subject of our inquiry is the logic of partial belief, and I do not think we can carry it far unless we have at least an approximate notion of what partial belief is, and how, if at all, it can be measured. It will not be very enlightening to be told that in such circumstances it would be rational to believe a proposition to the extent of $\frac{2}{3}$, unless we know what sort of a belief in it that means. We must therefore try to develop a purely psychological method of measuring belief. It is not enough to

[3] "Logical Constants", *Mind*, 1927.

measure probability; in order to apportion correctly our belief to the probability we must also be able to measure our belief.

It is a common view that belief and other psychological variables are not measurable, and if this is true our inquiry will be vain; and so will the whole theory of probability conceived as a logic of partial belief; for if the phrase 'a belief two-thirds of certainty' is meaningless, a calculus whose sole object is to enjoin such beliefs will be meaningless also. Therefore unless we are prepared to give up the whole thing as a bad job we are bound to hold that beliefs can to some extent be measured. If we were to follow the analogy of Mr Keynes' treatment of probabilities we should say that some beliefs were measurable and some not; but this does not seem to me likely to be a correct account of the matter: I do not see how we can sharply divide beliefs into those which have a position in the numerical scale and those which have not. But I think beliefs do differ in measurability in the following two ways. First, some beliefs can be measured more accurately than others; and, secondly, the measurement of beliefs is almost certainly an ambiguous process leading to a variable answer depending on how exactly the measurement is conducted. The degree of a belief is in this respect like the time interval between two events; before Einstein it was supposed that all the ordinary ways of measuring a time interval would lead to the same result if properly performed. Einstein showed that this was not the case; and time interval can no longer be regarded as an exact notion, but must be discarded in all precise investigations. Nevertheless, time interval and the Newtonian system are sufficiently accurate for many purposes and easier to apply.

I shall try to argue later that the degree of a belief is just like a time interval; it has no precise meaning unless we specify more exactly how it is to be measured. But for many purposes we can assume that the alternative ways of measuring it lead to the same result, although this is only approximately true. The resulting discrepancies are more glaring in connection with some beliefs than with others, and these therefore appear less measurable. Both these types of deficiency in measurability, due respectively to the difficulty in getting an exact enough measurement and to an important ambiguity in the definition of the measurement process, occur also in physics and so are not difficulties peculiar to our problem; what is peculiar is that it is difficult to form any idea of how the measurement is to be conducted, how a unit is to be obtained, and so on.

Let us then consider what is implied in the measurement of beliefs. A satisfactory system must in the first place assign to any belief a magnitude or degree having a definite position in an order of magnitudes; beliefs which are of the same degree as the same belief must be of the same degree as one another, and so on. Of course this cannot be accomplished without introducing a certain amount of hypothesis or fiction. Even in physics we cannot maintain that things that are equal to the same thing are equal to one another unless we take 'equal' not as meaning 'sensibly equal' but a fictitious or hypothetical relation. I do not want to discuss the metaphysics or epistemology of this process, but merely to remark that if it is allowable in physics it is allowable in psychology also. The logical simplicity characteristic of the relations dealt with in a science is never attained by nature alone without any admixture of fiction.

But to construct such an ordered series of degrees is not the whole of our task; we have also to assign numbers to these degrees in some intelligible manner. We can of course easily explain that we denote full belief by 1, full belief in the contradictory by 0, and equal beliefs in the proposition and its contradictory by $\frac{1}{2}$. But it is not so easy to say what is meant by a belief $\frac{2}{3}$ of certainty, or a belief in the proposition being twice as strong as that in its contradictory. This is the harder part of the task, but it is absolutely necessary; for we do calculate numerical probabilities, and if they are to correspond to degrees of belief we must discover some definite way of attaching numbers to degrees of belief. In physics we often attach numbers by discovering a physical process of addition[4]: the measure-numbers of lengths are not assigned arbitrarily subject only to the proviso that the greater length shall have the greater measure; we determine them further by deciding on a physical meaning for addition; the length got by putting together two given lengths must have for its measure the sum of their measures. A system of measurement in which there is nothing corresponding to this is immediately recognized as arbitrary, for instance Mohs' scale of hardness[5] in which 10 is arbitrarily assigned to diamond, the hardest known material, 9 to the next hardest, and so on. We have therefore to find a process of addition for degrees of belief, or some substitute for this which will be equally adequate to determine a numerical scale.

Such is our problem; how are we to solve it? There are, I think, two ways in which we can begin. We can, in the first place, suppose that the degree of a belief is something perceptible by its owner; for instance that beliefs differ in the intensity of a feeling by which they are accompanied, which might be called a belief-feeling or feeling of conviction, and that by the degree of belief we mean the intensity of this feeling. This view would be very inconvenient, for it is not easy to ascribe numbers to the intensities of feelings; but apart from this it seems to me observably false, for the beliefs which we hold most strongly are often accompanied by practically no feeling at all; no one feels strongly about things he takes for granted.

We are driven therefore to the second supposition that the degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it. This is a generalization of the well-known view, that the differentia of belief lies in its causal efficacy, which is discussed by Mr Russell in his *Analysis of Mind*. He there dismisses it for two reasons, one of which seems entirely to miss the point. He argues that in the course of trains of thought we believe many things which do not lead to action. This objection is however beside the mark, because it is not asserted that a belief is an idea which does actually lead to action, but one which would lead to action in suitable circumstances; just as a lump of arsenic is called poisonous not because it actually has killed or will kill anyone, but because it would kill anyone if he ate it. Mr Russell's second argument is, however, more formidable. He points out that it is not possible to suppose that beliefs differ from other ideas only in their effects, for if they were otherwise identical their effects

[4] See N. Campbell, *Physics The Elements* (1920), p.277.

[5] Ibid., p.271.

would be identical also. This is perfectly true, but it may still remain the case that the nature of the difference between the causes is entirely unknown or very vaguely known, and that what we want to talk about is the difference between the effects, which is readily observable and important.

As soon as we regard belief quantitatively, this seems to me the only view we can take of it. It could well be held that the difference between believing and not believing lies in the presence or absence of introspectible feelings. But when we seek to know what is the difference between believing more firmly and believing less firmly, we can no longer regard it as consisting in having more or less of certain observable feelings; at least I personally cannot recognize any such feelings. The difference seems to me to lie in how far we should act on these beliefs: this may depend on the degree of some feeling or feelings, but I do not know exactly what feelings and I do not see that it is indispensable that we should know. Just the same thing is found in physics; men found that a wire connecting plates of zinc and copper standing in acid deflected a magnetic needle in its neighbourhood. Accordingly as the needle was more or less deflected the wire was said to carry a larger or a smaller current. The nature of this 'current' could only be conjectured: what were observed and measured were simply its effects. It will no doubt be objected that we know how strongly we believe things, and that we can only know this if we can measure our belief by introspection. This does not seem to me necessarily true; in many cases, I think, our judgment about the strength of our belief is really about how we should act in hypothetical circumstances. It will be answered that we can only tell how we should act by observing the present belief-feeling which determines how we should act; but again I doubt the cogency of the argument. It is possible that what determines how we should act determines us also directly or indirectly to have a correct opinion as to how we should act, without its ever coming into consciousness.

Suppose, however, I am wrong about this and that we can decide by introspection the nature of belief, and measure its degree; still, I shall argue, the kind of measurement of belief with which probability is concerned is not this kind but is a measurement of belief *qua* basis of action. This can I think be shown in two ways. First, by considering the scale of probabilities between 0 and 1, and the sort of way we use it, we shall find that it is very appropriate to the measurement of belief as a basis of action, but in no way related to the measurement of an introspected feeling. For the units in terms of which such feelings or sensations are measured are always, I think, differences which are just perceptible: there is no other way of obtaining units. But I see no ground for supposing that the interval between a belief of degree $\frac{1}{3}$ and one of degree $\frac{1}{2}$ consists of as many just perceptible changes as does that between one of $\frac{2}{3}$ and one of $\frac{5}{6}$, or that a scale based on just perceptible differences would have any simple relation to the theory of probability. On the other hand the probability of $\frac{1}{3}$ is clearly related to the kind of belief which would lead to a bet of 2 to 1, and it will be shown below how to generalize this relation so as to apply to action in general. Secondly, the quantitative aspects of beliefs as the basis of action are evidently more important than the intensities of belief-feelings. The latter are no doubt interesting,

but may be very variable from individual to individual, and their practical interest is entirely due to their position as the hypothetical causes of beliefs *qua* bases of action.

It is possible that some one will say that the extent to which we should act on a belief in suitable circumstances is a hypothetical thing, and therefore not capable of measurement. But to say this is merely to reveal ignorance of the physical sciences which constantly deal with and measure hypothetical quantities; for instance, the electric intensity at a given point is the force which would act on a unit charge if it were placed at the point.

Let us now try to find a method of measuring beliefs as bases of possible actions. It is clear that we are concerned with dispositional rather than with actualized beliefs; that is to say, not with beliefs at the moment when we are thinking of them, but with beliefs like my belief that the earth is round, which I rarely think of, but which would guide my action in any case to which it was relevant.

The old-established way of measuring a person's belief is to propose a bet, and see what are the lowest odds which he will accept. This method I regard as fundamentally sound; but it suffers from being insufficiently general, and from being necessarily inexact. It is inexact partly because of the diminishing marginal utility of money, partly because the person may have a special eagerness or reluctance to bet, because he either enjoys or dislikes excitement or for any other reason, e.g. to make a book. The difficulty is like that of separating two different co-operating forces. Besides, the proposal of a bet may inevitably alter his state of opinion; just as we could not always measure electric intensity by actually introducing a charge and seeing what force it was subject to, because the introduction of the charge would change the distribution to be measured.

In order therefore to construct a theory of quantities of belief which shall be both general and more exact, I propose to take as a basis a general psychological theory, which is now universally discarded, but nevertheless comes, I think, fairly close to the truth in the sort of cases with which we are most concerned. I mean the theory that we act in the way we think most likely to realize the objects of our desires, so that a person's actions are completely determined by his desires and opinions. This theory cannot be made adequate to all the facts, but it seems to me a useful approximation to the truth particularly in the case of our self-conscious or professional life, and it is presupposed in a great deal of our thought. It is a simple theory and one which many psychologists would obviously like to preserve by introducing unconscious desires and unconscious opinions in order to bring it more into harmony with the facts. How far such fictions can achieve the required result I do not attempt to judge: I only claim for what follows approximate truth, or truth in relation to this artificial system of psychology, which like Newtonian mechanics can, I think, still be profitably used even though it is known to be false.

It must be observed that this theory is not to be identified with the psychology of the Utilitarians, in which pleasure had a dominating position. The theory I propose to adopt is that we seek things which we want, which may be our own or other people's pleasure, or anything else whatever, and our actions are such as we think most likely to realize these goods. But this is not a precise statement, for a precise statement of the theory can only be made after we have introduced the notion of quantity of belief.

Let us call the things a person ultimately desires 'goods', and let us at first assume that they are numerically measurable and additive. That is to say that if he prefers for its own sake an hour's swimming to an hour's reading, he will prefer two hours' swimming to one hour's swimming and one hour's reading. This is of course absurd in the given case but this may only be because swimming and reading are not ultimate goods, and because we cannot imagine a second hour's swimming precisely similar to the first, owing to fatigue, etc.

Let us begin by supposing that our subject has no doubts about anything, but certain opinions about all propositions. Then we can say that he will always choose the course of action which will lead in his opinion to the greatest sum of good.

It should be emphasized that in this essay good and bad are never to be understood in any ethical sense but simply as denoting that to which a given person feels desire and aversion.

The question then arises how we are to modify this simple system to take account of varying degrees of certainty in his beliefs. I suggest that we introduce as a law of psychology that his behaviour is governed by what is called the mathematical expectation; that is to say that, if p is a proposition about which he is doubtful, any goods or bads for whose realization p is in his view a necessary and sufficient condition enter into his calculations multiplied by the same fraction, which is called the 'degree of his belief in p'. We thus define degree of belief in a way which presupposes the use of the mathematical expectation.

We can put this in a different way. Suppose his degree of belief in p is $\frac{m}{n}$; then his action is such as he would choose it to be if he had to repeat it exactly n times, in m of which p was true, and in the others false. [Here it may be necessary to suppose that in each of the n times he had no memory of the previous ones.]

This can also be taken as a definition of the degree of belief, and can easily be seen to be equivalent to the previous definition. Let us give an instance of the sort of case which might occur. I am at a cross-roads and do not know the way; but I rather think one of the two ways is right. I propose therefore to go that way but keep my eyes open for someone to ask; if now I see someone half a mile away over the fields, whether I turn aside to ask him will depend on the relative inconvenience of going out of my way to cross the fields or of continuing on the wrong road if it is the wrong road. But it will also depend on how confident I am that I am right; and clearly the more confident I am of this the less distance I should be willing to go from the road to check my opinion. I propose therefore to use the distance I would be prepared to go to ask, as a measure of the confidence of my opinion; and what I have said above explains how this is to be done. We can set it out as follows: suppose the disadvantage of going x yards to ask is $f(x)$, the advantage of arriving at the right destination is r, that of arriving at the wrong one w. Then if I should just be willing to go a distance d to ask, the degree of my belief that I am on the right road is given by

$$p = 1 - \frac{f(d)}{r - w}.$$

For such an action is one it would just pay me to take, if I had to act in the same way n times, in np of which I was on the right way but in the others not.

For the total good resulting from not asking each time

$$= npr + n(1-p)w$$
$$= nw + np(r-w)$$

that resulting from asking at distance x each time

$$= nr - nf(x). \qquad \text{[I now always go right.]}$$

This is greater than the preceding expression, provided

$$f(x) < (r-w)(1-p),$$

$\therefore$ the critical distance d is connected with p, the degree of belief, by the relation $f(d) = (r-w)(1-p)$

$$\text{or} \quad p = 1 - \frac{f(d)}{r-w} \quad \text{as asserted above.}$$

It is easy to see that this way of measuring belief gives results agreeing with ordinary ideas; at any rate to the extent that full belief is denoted by 1, full belief in the contradictory by 0, and equal belief in the two by $\frac{1}{2}$. Further, it allows validity to betting as means of measuring beliefs. By proposing a bet on p we give the subject a possible course of action from which so much extra good will result to him if p is true and so much extra bad if p is false. Supposing, the bet to be in goods and bads instead of in money, he will take a bet at any better odds than those corresponding to his state of belief; in fact his state of belief is measured by the odds he will just take; but this is vitiated, as already explained, by love or hatred of excitement, and by the fact that the bet is in money and not in goods and bads. Since it is universally agreed that money has a diminishing marginal utility, if money bets are to be used, it is evident that they should be for as small stakes as possible. But then again the measurement is spoiled by introducing the new factor of reluctance to bother about trifles.

Let us now discard the assumption that goods are additive and immediately measurable, and try to work out a system with as few assumptions as possible. To begin with we shall suppose, as before, that our subject has certain beliefs about everything; then he will act so that what he believes to be the total consequences of his action will be the best possible. If then we had the power of the Almighty, and could persuade our subject of our power, we could, by offering him options, discover how he placed in order of merit all possible courses of the world. In this way all possible worlds would be put in an order of value, but we should have no definite way of representing them by numbers. There would be no meaning in the assertion that the difference in value between α and β was equal to that between γ and δ. [Here and elsewhere we use Greek letters to represent the different possible totalities of events between which our subject chooses – the ultimate organic unities.]

Suppose next that the subject is capable of doubt; then we could test his degree of belief in different propositions by making him offers of the following kind. Would you rather have world α in any event; or world β if *p* is true, and world γ if *p* is false? If, then, he were certain that *p* was true, simply compare α and β and choose between them as if no conditions were attached; but if he were doubtful his choice would not be decided so simply. I propose to lay down axioms and definitions concerning the principles governing choices of this kind. This is, of course, a very schematic version of the situation in real life, but it is, I think, easier to consider it in this form.

There is first a difficulty which must be dealt with; the propositions like *p* in the above case which are used as conditions in the options offered may be such that their truth or falsity is an object of desire to the subject. This will be found to complicate the problem, and we have to assume that there are propositions for which this is not the case, which we shall call ethically neutral. More precisely an atomic proposition *p* is called ethically neutral if two possible worlds differing only in regard to the truth of *p* are always of equal value; and a non-atomic proposition *p* is called ethically neutral if all its atomic truth-arguments[6] are ethically neutral.

We begin by defining belief of degree $\frac{1}{2}$ in an ethically neutral proposition. The subject is said to have belief of degree $\frac{1}{2}$ in such a proposition *p* if he has no preference between the options (1) α if *p* is true, β if *p* is false, and (2) α if *p* is false, β if *p* is true, but has a preference between α and β simply. We suppose by an axiom that if this is true of any one pair α, β, it is true of all such pairs.[7] This comes roughly to defining belief of degree $\frac{1}{2}$ as such a degree of belief as leads to indifference between betting one way and betting the other for the same stakes.

Belief of degree $\frac{1}{2}$ as thus defined can be used to measure values numerically in the following way. We have to explain what is meant by the difference in value between α and β being equal to that between γ and δ; and we define this to mean that, if *p* is an ethically neutral proposition believed to degree $\frac{1}{2}$, the subject has no preference between the options (1) α if *p* is true, δ if *p* is false, and (2) β if *p* is true, γ if *p* is false.

This definition can form the basis of a system of measuring values in the following way:–

Let us call any set of all worlds equally preferable to a given world a value: we suppose that if world α is preferable to β any world with the same value as α is preferable to any world with the same value as β and shall say that the value of α is greater than that of β. This relation 'greater than' orders values in a series. We shall use α henceforth both for the world and its value.

[6] I assume here Wittgenstein's theory of propositions; it would probably be possible to give an equivalent definition in terms of any other theory.

[7] α and β must be supposed so far undefined as to be compatible with both *p* and not-*p*.

Axioms

(1) There is an ethically neutral proposition p believed to degree $\frac{1}{2}$.
(2) If p, q are such propositions and the option

α if p, δ if not-p is equivalent to β if p, γ if not-p
then α if q, δ if not-q is equivalent to β if q, γ if not-q.
Def. In the above case we say $\alpha\beta = \gamma\delta$.

Theorems If $\alpha\beta = \gamma\delta$,
then $\beta\alpha = \delta\gamma$, $\alpha\gamma = \beta\delta$, $\gamma\alpha = \delta\beta$.

(2a) If $\alpha\beta = \gamma\delta$, then $\alpha > \beta$ is equivalent to $\gamma > \delta$
and $\alpha = \beta$ is equivalent to $\gamma = \delta$.

(3) If option A is equivalent to option B and B to C, then A to C.

Theorem If $\alpha\beta = \gamma\delta$ and $\beta\eta = \zeta\gamma$,
then $\alpha\eta = \zeta\delta$.

(4) If $\alpha\beta = \gamma\delta$, $\gamma\delta = \eta\zeta$, then $\alpha\beta = \eta\zeta$.
(5) (α, β, γ). E!(ιx) $(\alpha x = \beta\gamma)$
(6) (α, β). E!(ιx) $(\alpha x = x\beta)$
(7) Axiom of continuity: – Any progression has a limit (ordinal).
(8) Axiom of Archimedes.

These axioms enable the values to be correlated one-one with real numbers so that if α^1 corresponds to α, etc.

$$\alpha\beta = \gamma\delta. \equiv .\alpha^1 - \beta^1 = \gamma^1 - \delta^1.$$

Henceforth we use α for the correlated real number α^1 also.

Having thus defined a way of measuring value we can now derive a way of measuring belief in general. If the option of α for certain is indifferent with that of β if p is true and γ if p is false,[8] we can define the subject's degree of belief in p as the ratio of the difference between α and γ to that between β and γ; which we must suppose the same for all α's, β's and γ's that satisfy the conditions. This amounts roughly to defining the degree of belief in p by the odds at which the subject would bet on p, the bet being conducted in terms of differences of value as defined. The definition only applies to partial belief and does not include certain beliefs; for belief of degree 1 in p, α for certain is indifferent with α if p and any β if not-p.

We are also able to define a very useful new idea – the 'degree of belief in p given q'. This does not mean the degree of belief in 'If p then q', or that in 'p entails q', or that which the subject would have in p if he knew q, or that which he ought to have.

[8] Here β must include the truth of p, γ its falsity; p need no longer be ethically neutral. But we have to assume that there is a wolrd with any assigned value in which p is true, and one in which p is false.

It roughly expresses the odds at which he would now bet on p, the bet only to be valid if q is true. Such conditional bets were often made in the eighteenth century.

The degree of belief in p given q is measured thus. Suppose the subject indifferent between the options (1) α if q true, β if q false, (2) γ if p true and q true, δ if p false and q true, β if q false. Then the degree of his belief in p given q is the ratio of the difference between α and δ to that between γ and δ, which we must suppose the same for any α, β, γ, δ which satisfy the given conditions. This is not the same as the degree to which he would believe p, if he believed q for certain; for knowledge of q might for psychological reasons profoundly alter his whole system of beliefs.

Each of our definitions has been accompanied by an axiom of consistency, and in so far as this is false, the notion of the corresponding degree of belief becomes invalid. This bears some analogy to the situation in regard to simultaneity discussed above.

I have not worked out the mathematical logic of this in detail, because this would, I think, be rather like working out to seven places of decimals a result only valid to two. My logic cannot be regarded as giving more than the sort of way it might work.

From these definitions and axioms it is possible to prove the fundamental laws of probable belief (degrees of belief lie between 0 and 1):

(1) Degree of belief in p + degree of belief in $\overline{p} = 1$
(2) Degree of belief in p given q + degree of belief in $\overline{p}$ given $q = 1$.
(3) Degree of belief in (p and q) = degree of belief in $p \times$ degree of belief in q given p.
(4) Degree of belief in (p and q) + degree of belief in (p and $\overline{q}$) = degree of belief in p.

The first two are immediate. (3) is proved as follows.

Let degree of belief in $p = x$, that in q given $p = y$.

Then ξ for certain $\equiv \xi + (1-x)t$ if p true, $\xi - xt$ if p false for any t.

$$\xi + (1-x)\,t \text{ if } p \text{ true} \equiv$$

$$\begin{cases} \xi + (1-x)\,t + (1-y)\,u \text{ if '}p \text{ and } q\text{' true,} \\ \xi + (1-x)\,t - yu \text{ if } p \text{ true } q \text{ false;} \quad \text{for any } u. \end{cases}$$

Choose u so that $\xi + (1-x)\,t - yu = \xi - xt$,

$$\text{i.e. let } u = t/y\,(y \neq 0)$$

Then ξ for certain $\equiv$

$$\begin{cases} \xi + (1-x)\,t + (1-y)\,t/y \text{ if } p \text{ and } q \text{ true} \\ \xi - xt \text{ otherwise,} \end{cases}$$

$\therefore$ degree of belief in 'p and q' $= \frac{xt}{t+(1-y)t/y} = xy.$ $(t \neq 0)$
If $y = 0$, take $t = 0$.

$$\text{Then } \xi \text{ for certain} \equiv \xi \text{ if } p \text{ true}, \quad \xi \text{ if } p \text{ false}$$
$$\equiv \xi + u \text{ if } p \text{ true}, q \text{ true}; \xi \text{ if } p \text{ false}, q \text{ false}, \xi \text{ if } p \text{ false}$$
$$\equiv \xi + u, pq \text{ true}; \xi, pq \text{ false}$$

$\therefore$ degree of belief in $pq = 0$.

(4) follows from (2), (3) as follows: –

Degree of belief in pq = that in $p \times$ that in q given p, by (3). Similarly degree of belief in $p\overline{q}$ = that in $p \times$ that in q given p

$\therefore$ sum = degree of belief in p, by (2).

These are the laws of probability, which we have proved to be necessarily true of any consistent set of degrees of belief. Any definite set of degrees of belief which broke them would be inconsistent in the sense that it violated the laws of preference between options, such as that preferability is a transitive asymmetrical relation, and that if α is preferable to β, β for certain cannot be preferable to α if p, β if not-p. If anyone's mental condition violated these laws, his choice would depend on the precise form in which the options were offered him, which would be absurd. He could have a book made against him by a cunning better and would then stand to lose in any event.

We find, therefore, that a precise account of the nature of partial belief reveals that the laws of probability are laws of consistency, an extension to partial beliefs of formal logic, the logic of consistency. They do not depend for their meaning on any degree of belief in a proposition being uniquely determined as the rational one; they merely distinguish those sets of beliefs which obey them as consistent ones.

Having any definite degree of belief implies a certain measure of consistency, namely willingness to bet on a given proposition at the same odds for any stake, the stakes being measured in terms of ultimate values. Having degrees of belief obeying the laws of probability implies a further measure of consistency, namely such a consistency between the odds acceptable on different propositions as shall prevent a book being made against you.

Some concluding remarks on this section may not be out of place. First, it is based fundamentally on betting, but this will not seem unreasonable when it is seen that all our lives we are in a sense betting. Whenever we go to the station we are betting that a train will really run, and if we had not a sufficient degree of belief in this we should decline the bet and stay at home. The options God gives us are always conditional on our guessing whether a certain proposition is true. Secondly, it is based throughout on the idea of mathematical expectation; the dissatisfaction often felt with this idea is due mainly to the inaccurate measurement of goods. Clearly mathematical expectations in terms of money are not proper guides to conduct. It should be remembered, in judging my system, that in it value is actually defined by means of mathematical expectation in the case of beliefs of degree $\frac{1}{2}$, and so may be expected to be scaled suitably for the valid application in the case of other degrees of belief also.

Thirdly, nothing has been said about degrees of belief when the number of alternatives is infinite. About this I have nothing useful to say, except that I doubt if the mind is capable of contemplating more than a finite number of alternatives. It can consider questions to which an infinite number of answers are possible, but in order to consider the answers it must lump them into a finite number of groups. The difficulty becomes practically relevant when discussing induction, but even then there seems to me no need to introduce it. We can discuss whether past experience gives a high probability to the sun's rising to-morrow without bothering about what probability it gives to the sun's rising each morning for evermore. For this reason I cannot but feel that Mr Ritchie's discussion of the problem[9] is unsatisfactory; it is true that we can agree that inductive generalizations need have no finite probability, but particular expectations entertained on inductive grounds undoubtedly do have a high numerical probability in the minds of all of us. We all are more certain that the sun will rise to-morrow than that I shall not throw 12 with two dice first time, i.e. we have a belief of higher degree than $\frac{35}{36}$ in it. If induction ever needs a logical justification it is in connection with the probability of an event like this.

The Logic of Consistency

We may agree that in some sense it is the business of logic to tell us what we ought to think; but the interpretation of this statement raises considerable difficulties. It may be said that we ought to think what is true, but in that sense we are told what to think by the whole of science and not merely by logic. Nor, in this sense, can any justification be found for partial belief; the ideally best thing is that we should have beliefs of degree 1 in all true propositions and beliefs of degree 0 in all false propositions. But this is too high a standard to expect of mortal men, and we must agree that some degree of doubt or even of error may be humanly speaking justified.

Many logicians, I suppose, would accept as an account of their science the opening words of Mr Keynes' *Treatise on Probability*: "Part of our knowledge we obtain direct; and part by argument. The Theory of Probability is concerned with that part which we obtain by argument, and it treats of the different degrees in which the results so obtained are conclusive or inconclusive." Where Mr Keynes says 'the Theory of Probability', others would say Logic. It is held, that is to say, that our opinions can be divided into those we hold immediately as a result of perception or

[9] A. D. Ritchie, "Induction and Probability." *Mind*, 1926. p. 318. 'The conclusion of the foregoing discussion may be simply put. If the problem of induction be stated to be "How can inductive generalizations acquire a large numerical probability?" then this is a pseudo-problem, because the answer is "They cannot". This answer is not, however, a denial of the validity of induction but is a direct consequence of the nature of probability. It still leaves untouched the real problem of induction which is "How can the probability of an induction be increased?" and it leaves standing the whole of Keynes' discussion on this point.'

memory, and those which we derive from the former by argument. It is the business of Logic to accept the former class and criticize merely the derivation of the second class from them.

Logic as the science of argument and inference is traditionally and rightly divided into deductive and inductive; but the difference and relation between these two divisions of the subject can be conceived in extremely different ways. According to Mr Keynes valid deductive and inductive arguments are fundamentally alike; both are justified by logical relations between premiss and conclusion which differ only in degree. This position, as I have already explained, I cannot accept. I do not see what these inconclusive logical relations can be or how they can justify partial beliefs. In the case of conclusive logical arguments I can accept the account of their validity which has been given by many authorities, and can be found substantially the same in Kant, De Morgan, Peirce and Wittgenstein. All these authors agree that the conclusion of a formally valid argument is contained in its premisses; that to deny the conclusion while accepting the premisses would be self-contradictory; that a formal deduction does not increase our knowledge, but only brings out clearly what we already know in another form; and that we are bound to accept its validity on pain of being inconsistent with ourselves. The logical relation which justifies the inference is that the sense or import of the conclusion is contained in that of the premisses.

But in the case of an inductive argument this does not happen in the least; it is impossible to represent it as resembling a deductive argument and merely weaker in degree; it is absurd to say that the sense of the conclusion is partially contained in that of the premisses. We could accept the premisses and utterly reject the conclusion without any sort of inconsistency or contradiction.

It seems to me, therefore, that we can divide arguments into two radically different kinds, which we can distinguish in the words of Peirce as (1) 'explicative, analytic, or deductive' and (2) 'amplifiative, synthetic, or (loosely speaking) inductive'.[10] Arguments of the second type are from an important point of view much closer to memories and perceptions than to deductive arguments. We can regard perception, memory and induction as the three fundamental ways of acquiring knowledge; deduction on the other hand is merely a method of arranging our knowledge and eliminating inconsistencies or contradictions.

Logic must then fall very definitely into two parts: (excluding analytic logic, the theory of terms and propositions) we have the lesser logic, which is the logic of consistency, or formal logic; and the larger logic, which is the logic of discovery, or inductive logic.

What we have now to observe is that this distinction in no way coincides with the distinction between certain and partial beliefs; we have seen that there is a theory of consistency in partial beliefs just as much as of consistency in certain beliefs, although for various reasons the former is not so important as the latter. The theory of probability is in fact a generalization of formal logic; but in the process

[10]C.S. Peirce *Change Love and Logic*, p. 92.

of generalization one of the most important aspects of formal logic is destroyed. If p and $\overline{q}$ are inconsistent so that q follows logically from p, that p implies q is what is called by Wittgenstein a 'tautology' and can be regarded as a degenerate case of a true proposition not involving the idea of consistency. This enables us to regard (not altogether correctly) formal logic including mathematics as an objective science consisting of objectively necessary propositions. It thus gives us not merely the ἀνάγκη λέγειν, that if we assert p we are bound in consistency to assert q also, but also the ἀνάγκη εἶναι, that is p is true, so must q be. But when we extend formal logic to include partial beliefs this direct objective interpretation is lost; if we believe pq to the extent of $\frac{1}{3}$, and $p\overline{q}$ to the extent of $\frac{1}{3}$, we are bound in consistency to believe $\overline{p}$ also to the extent of $\frac{1}{3}$. This is the ἀνάγκη λέγειν; but we cannot say that if pq is $\frac{1}{3}$ true and $p\overline{q}$ $\frac{1}{3}$ true, $\overline{p}$ also must be $\frac{1}{3}$ true, for such a statement would be sheer nonsense. There is no corresponding ἀνάγκη εἶναι. Hence, unlike the calculus of consistent full belief, the calculus of objective partial belief cannot be immediately interpreted as a body of objective tautology.

This is, however, possible in a roundabout way; we saw at the beginning of this essay that the calculus of probabilities could be interpreted in terms of class-ratios; we have now found that it can also be interpreted as a calculus of consistent partial belief. It is natural, therefore, that we should expect some intimate connection between these two interpretations, some explanation of the possibility of applying the same mathematical calculus to two such different sets of phenomena. Nor is an explanation difficult to find; there are many connections between partial beliefs and frequencies. For instance, experienced frequencies often lead to corresponding partial beliefs, and partial beliefs lead to the expectation of corresponding frequencies in accordance with Bernouilli's Theorem. But neither of these is exactly the connection we want; a partial belief cannot in general be connected uniquely with any actual frequency, for the connection is always made by taking the proposition in question as an instance of a propositional function. What propositional function we choose is to some extent arbitrary and the corresponding frequency will vary considerably with our choice. . The pretensions of some exponents of the frequency theory that partial belief means full belief in a frequency proposition cannot be sustained. But we found that the very idea of partial belief involves reference to a hypothetical or ideal frequency; supposing goods to be additive, belief of degree $\frac{m}{n}$ is the sort of belief which leads to the action which would be best if repeated n times in m of which the proposition is true; or we can say more briefly that it is the kind of belief most appropriate to a number of hypothetical occasions otherwise identical in a proportion $\frac{m}{n}$ of which the proposition in question is true. It is this connection between partial belief and frequency which enables us to use the calculus of frequencies as a calculus of consistent partial belief. And in a sense we may say that the two interpretations are the objective and subjective aspects of the same inner meaning, just as formal logic can be interpreted objectively as a body of tautology and subjectively as the laws of consistent thought.

We shall, I think, find that this view of the calculus of probability removes various difficulties that have hitherto been found perplexing. In the first place it

gives us a clear justification for the axioms of the calculus, which on such a system as Mr Keynes' is entirely wanting. For now it is easily seen that if partial beliefs are consistent they will obey these axioms, but it is utterly obscure why Mr Keynes' mysterious logical relations should obey them.[11] We should be so curiously ignorant of the instances of these relations, and so curiously knowledgeable about their general laws.

Secondly, the Principle of Indifference can now be altogether dispensed with; we do not regard it as belonging to formal logic to say what should be a man's expectation of drawing a white or a black ball from an urn; his original expectations may within the limits of consistency be any he likes; all we have to point out is that if he has certain expectations he is bound in consistency to have certain others. This is simply bringing probability into line with ordinary formal logic, which does not criticize premisses but merely declares that certain conclusions are the only ones consistent with them. To be able to turn the Principle of Indifference out of formal logic is a great advantage; for it is fairly clearly impossible to lay down purely logical conditions for its validity, as is attempted by Mr Keynes. I do not want to discuss this question in detail, because it leads to hair-splitting and arbitrary distinctions which could be discussed for ever. But anyone who tries to decide by Mr Keynes' methods what are the proper alternatives to regard as equally probable in molecular mechanics, e.g. in Gibbs' phase-space, will soon be convinced that it is a matter of physics rather than pure logic. By using the multiplication formula, as it is used in inverse probability, we can on Mr Keynes' theory reduce all probabilities to quotients of *a priori* probabilities; it is therefore in regard to these latter that the Principle of Indifference is of primary importance; but here the question is obviously not one of formal logic. How can we on merely logical grounds divide the spectrum into equally probable bands?

A third difficulty which is removed by our theory is the one which is presented to Mr Keynes' theory by the following case. I think I perceive or remember something but am not sure; this would seem to give me some ground for believing it, contrary to Mr Keynes' theory, by which the degree belief in it which it would be rational for me to have is that given by the probability relation between the proposition in question and the things I know for certain. He cannot justify a probable belief founded not on argument but on direct inspection. In our view there would be nothing contrary to formal logic in such a belief; whether it would be reasonable would depend on what I have called the larger logic which will be the subject of the next section; we shall there see that there is no objection to such a possibility, with which Mr Keynes' method of justifying probable belief solely by relation to certain knowledge is quite unable to cope.

[11] It appears in Mr Keynes' system as if the principal axioms – the laws of addition and multiplication – were nothing but definitions. This is merely a logical mistake; his definitions are formally invalid unless corresponding axioms are presupposed. Thus his definition of multiplication presupposes the law that if the probability of *a* given *bh is* equal to that of *c* given *dh,* and the probability of *b* given *h* is equal to that of *d* given *h*, then will the probabilities of *ab* given *h* and of *cd* given *h* be equal.

The Logic of Truth

The validity of the distinction between the logic of consistency and the logic of truth has been often disputed; it has been contended on the one hand that logical consistency is only a kind of factual consistency; that if a belief in p is inconsistent with one in q, that simply means that p and q are not both true, and that this is a necessary or logical fact. I believe myself that this difficulty can be met by Wittgenstein's theory of tautology, according to which if a belief in p is inconsistent with one in q, that p and q are not both true is not a fact but a tautology. But I do not propose to discuss this question further here.

From the other side it is contended that formal logic or the logic of consistency is the whole of logic, and inductive logic either nonsense or part of natural science. This contention, which would I suppose be made by Wittgenstein, I feel more difficulty in meeting. But I think it would be a pity, out of deference to authority, to give up trying to say anything useful about induction.

Let us therefore go back to the general conception of logic as the science of rational thought. We found that the most generally accepted parts of logic, namely, formal logic, mathematics and the calculus of probabilities, are all concerned simply to ensure that our beliefs are not self-contradictory. We put before ourselves the standard of consistency and construct these elaborate rules to ensure its observance. But this is obviously not enough; we want our beliefs to be consistent not merely with one another but also with the facts[12]: nor is it even clear that consistency is always advantageous; it may well be better to be sometimes right than never right. Nor when we wish to be consistent are we always able to be: there are mathematical propositions whose truth or falsity cannot as yet be decided. Yet it may humanly speaking be right to entertain a certain degree of belief in them on inductive or other grounds: a logic which proposes to justify such a degree of belief must be prepared actually to go against formal logic; for to a formal truth formal logic can only assign a belief of degree 1. We could prove in Mr Keynes' system that its probability is 1 on any evidence. This point seems to me to show particularly clearly that human logic or the logic of truth, which tells men how they should think, is not merely independent of but sometimes actually incompatible with formal logic.

In spite of this nearly all philosophical thought about human logic and especially induction has tried to reduce it in some way to formal logic. Not that it is supposed, except by a very few, that consistency will of itself lead to truth; but consistency combined with observation and memory is frequently credited with this power.

Since an observation changes (in degree at least) my opinion about the fact observed, some of my degrees of belief after the observation are necessarily

[12]Cf. Kant: 'Denn obgleich eine Erkenntnis der logischen Form völlig gemäss sein möchte, dass ist sich selbst nicht widerspräche, so kann sie doch noch immer dem Gegenstande widersprechen.' *Kritik der reinen Vernunft*, First Edition. p. 59.

inconsistent with those I had before. We have therefore to explain how exactly the observation should modify my degrees of belief; obviously if p is the fact observed, my degree of belief in q after the observation should be equal to my degree of belief in q given p before, or by the multiplication law to the quotient of my degree of belief in pq by my degree of belief in p. When my degrees of belief change in this way we can say that they have been changed consistently by my observation.

By using this definition, or on Mr Keynes' system simply by using the multiplication law, we can take my present degrees of belief, and by considering the totality of my observations, discover from what initial degrees of belief my present ones would have arisen by this process of consistent change. My present degrees of belief can then be considered logically justified if the corresponding initial degrees of belief are logically justified. But to ask what initial degrees of belief are justified, or in Mr Keynes' system what are the absolutely *a priori* probabilities, seems to me a meaningless question; and even if it had a meaning I do not see how it could be answered.

If we actually applied this process to a human being, found out, that is to say, on what *a prior* probabilities his present opinions could be based, we should obviously find them to be ones determined by natural selection, with a general tendency to give a higher probability to the simpler alternatives. But, as I say, I cannot see what could be meant by asking whether these degrees of belief were logically justified. Obviously the best thing would be to know for certain in advance what was true and what false, and therefore if any one system of initial beliefs is to receive the philosopher's approbation it should be this one. But clearly this would not be accepted by thinkers of the school I am criticising. Another alternative is to apportion initial probabilities on the purely formal system expounded by Wittgenstein, but as this gives no justification for induction it cannot give us the human logic which we are looking for.

Let us therefore try to get an idea of a human logic which shall not attempt to be reducible to formal logic. Logic, we may agree, is concerned not with what men actually believe, but what they ought to believe, or what it would be reasonable to believe. What then, we must ask, is meant by saying that it is reasonable for a man to have such and such a degree of belief in a proposition? Let us consider possible alternatives.

First, it sometimes means something explicable in terms of formal logic: this possibility for reasons already explained we may dismiss. Secondly, it sometimes means simply that were I in his place (and not e.g. drunk) I should have such a degree of belief. Thirdly, it sometimes means that if his mind worked according to certain rules, which we may roughly call 'scientific method', he would have such a degree of belief. But fourthly it need mean none of these things for men have not always believed in scientific method, and just as we ask 'But am I necessarily reasonable,' we can also ask 'But is the scientist necessarily reasonable?' In this ultimate meaning it seems to me that we can identify reasonable opinion with the opinion of an ideal person in similar circumstances. What, however, would this ideal person's opinion be? As has previously been remarked, the highest ideal would be

always to have a true opinion and be certain of it; but this ideal is more suited to God than to man.[13]

We have therefore to consider the human mind and what is the most we can ask of it.[14] The human mind works essentially according to general rules or habits; a process of thought not proceeding according to some rule would simply be a random sequence of ideas; whenever we infer *A* from *B* we do so in virtue of some relation between them. We can therefore state the problem of the ideal as "What habits in a general sense would it be best for the human mind to have?" This is a large and vague question which could hardly be answered unless the possibilities were first limited by a fairly definite conception of human nature. We could imagine some very useful habits unlike those possessed by any men. [It must be explained that I use habit in the most general possible sense to mean simply rule or law of behaviour, including instinct: I do not wish to distinguish acquired rules or habits in the narrow sense from innate rules or instincts, but propose to call them all habits alike.] A completely general criticism of the human mind is therefore bound to be vague and futile, but something useful can be said if we limit the subject in the following way.

Let us take a habit of forming opinion in a certain way; e.g. the habit of proceeding from the opinion that a toadstool is yellow to the opinion that it is unwholesome. Then we can accept the fact that the person has a habit of this sort, and ask merely what degree of opinion that the toadstool is unwholesome it would be best for him to entertain when he sees it; i.e. granting that he is going to think always in the same way about all yellow toadstools, we can ask what degree of confidence it would be best for him to have that they are unwholesome. And the answer is that it will in general be best for his degree of belief that a yellow toadstool is unwholesome to be equal to the proportion of yellow toadstools which are in fact

[13] [Earlier draft of matter of preceding paragraph in some ways better. – F.P.R.

What is meant by saying that a degree of belief is reasonable? First and often that it is what I should entertain if I had the opinions of the person in question at the time but was otherwise as I am now, e.g. not drunk. But sometimes we go beyond this and ask: 'Am I reasonable?' This may mean, do I conform to certain enumerable standards which we call scientific method, and which we value on account of those who practise them and the success they achieve. In this sense to be reasonable means to think like a scientist, or to be guided only by ratiocination and induction or something of the sort (i.e. reasonable means reflective). Thirdly, we may go to the root of why we admire the scientist and criticize not primarily an individual opinion but a mental habit as being conducive or otherwise to the discovery of truth or to entertaining such degrees of belief as will be most useful. (To include habits of doubt or partial belief.) Then we can criticize an opinion according to the habit which produced it. This is clearly right because it all depends on this habit; it would not be reasonable to get the right conclusion to a syllogism by remembering vaguely that you leave out a term which is common to both premisses.

We use reasonable in sense 1 when we say of an argument of a scientist this does not seem to me reasonable; in sense 2 when we *contrast* reason and superstition or instinct; in sense 3 when we *estimate* the value of new methods of thought such as soothsaying.].

[14] What follows to the end of the section is almost entirely based on the writings of C. S. Peirce. [Especially his "Illustrations of the Logic of Science", *Popular Science Monthly*, 1877 and 1878, reprinted in *Chance Love and Logic* (1923).].

unwholesome. (This follows from the meaning of degree of belief.) This conclusion is necessarily vague in regard to the spatio-temporal range of toadstools which it includes, but hardly vaguer than the question which it answers. (Cf. density at a point of gas composed of molecules.)

Let us put it in another way: whenever I make an inference, I do so according to some rule or habit. An inference is not completely given when we are given the premiss and conclusion; we require also to be given the relation between them in virtue of which the inference is made. The mind works by general laws; therefore if it infers q from p, this will generally be because q is an instance of a function φx and p the corresponding instance of a function ψx such that the mind would always infer φx from ψx. When therefore we criticize not opinions but the processes by which they are formed, the rule of the inference determines for us a range to which the frequency theory can be applied. The rule of the inference may be narrow, as when seeing lightning I expect thunder, or wide, as when considering 99 instances of a generalization which I have observed to be true I conclude that the 100th is true also. In the first case the habit which determines the process is 'After lightning expect thunder'; the degree of expectation which it would be best for this habit to produce is equal to the proportion of cases of lightning which are actually followed by thunder. In the second case the habit is the more general one of inferring from 99 observed instances of a certain sort of generalization that the 100th instance is true also; the degree of belief it would be best for this habit to produce is equal to the proportion of all cases of 99 instances of a generalization being true, in which the 100th is true also.

Thus given a single opinion, we can only praise or blame it on the ground of truth or falsity: given a habit of a certain form, we can praise or blame it accordingly as the degree of belief it produces is near or far from the actual proportion in which the habit leads to truth. We can then praise or blame opinions derivatively from our praise or blame of the habits that produce them.

This account can be applied not only to habits of inference but also to habits of observation and memory; when we have a certain feeling in connection with an image we think the image represents something which actually happened to us, but we may not be sure about it; the degree of direct confidence in our memory varies. If we ask what is the best degree of confidence to place in a certain specific memory feeling, the answer must depend on how often when that feeling occurs the event whose image it attaches to has actually taken place.

Among the habits of the human mind a position of peculiar importance is occupied by induction. Since the time of Hume a great deal has been written about the justification for inductive inference. Hume showed that it could not be reduced to deductive inference or justified by formal logic. So far as it goes his demonstration seems to me final; and the suggestion of Mr Keynes that it can be got round by regarding induction as a form of probable inference cannot in my view be maintained. But to suppose that the situation which results from this is a scandal to philosophy is, I think, a mistake.

We are all convinced by inductive arguments, and our conviction is reasonable because the world is so constituted that inductive arguments lead on the whole to true opinions. We are not, therefore, able to help trusting induction, nor if we could

help it do we see any reason why we should, because we believe it to be a reliable process. It is true that if any one has not the habit of induction, we cannot prove to him that he is wrong; but there is nothing peculiar in that. If a man doubts his memory or his perception we cannot prove to him that they are trustworthy; to ask for such a thing to be proved is to cry for the moon, and the same is true of induction. It is one of the ultimate sources of knowledge just as memory is: no one regards it as a scandal to philosophy that there is no proof that the world did not begin two minutes ago and that all our memories are not illusory.

We all agree that a man who did not make inductions would be unreasonable: the question is only what this means. In my view it does not mean that the man would in any way sin against formal logic or formal probability; but that he had not got a very useful habit, without which he would be very much worse off, in the sense of being much less likely[15] to have true opinions.

This is a kind of pragmatism: we judge mental habits by whether they work, i.e. whether the opinions they lead to are for the most part true, or more often true than those which alternative habits would lead to.

Induction is such a useful habit, and so to adopt it is reasonable. All that philosophy can do is to analyse it, determine the degree of its utility, and find on what characteristics of nature this depends. An indispensable means for investigating these problems is induction itself, without which we should be helpless. In this circle lies nothing vicious. It is only through memory that we can determine the degree of accuracy of memory; for if we make experiments to determine this effect, they will be useless unless we remember them.

Let us consider in the light of the preceding discussion what sort of subject is inductive or human logic – the logic of truth. Its business is to consider methods of thought, and discover what degree of confidence should be placed in them, i.e. in what proportion of cases they lead to truth. In this investigation it can only be distinguished from the natural sciences by the greater generality of its problems. It has to consider the relative validity of different types of scientific procedure, such as the search for a causal law by Mill's Methods, and the modern mathematical methods like the *a priori* arguments used in discovering the Theory of Relativity. The proper plan of such a subject is to be found in Mill[16]; I do not mean the details of his Methods or even his use of the Law of Causality. But his way of treating the subject as a body of inductions about inductions, the Law of Causality governing lesser laws and being itself proved by induction by simple enumeration. The different scientific methods that can be used are in the last resort judged by induction by simple enumeration; we choose the simplest law that fits the facts, but unless we found that laws so obtained also fitted facts other than those they were made to fit, we should discard this procedure for some other.

[15] 'Likely' here simply means that I am not sure of this, but only have a certain degree of belief in it.

[16] Cf. also the account of 'general rules' in the Chapter 'Of Unphilosophical Probability' in Hume's *Treatise*.

Chapter 4
Probable Knowledge

Richard C. Jeffrey

The central problem of epistemology is often taken to be that of explaining how we can know what we do, but the content of this problem changes from age to age with the scope of what we take ourselves to know; and philosophers who are impressed with this flux sometimes set themselves the problem of explaining how we can get along, knowing as little as we do. For knowledge is sure, and there seems to be little we can be sure of outside logic and mathematics and truths related immediately to experience. It is as if there were some propositions – that this paper is white, that two and two are four – on which we have a firm grip, while the rest, including most of the theses of science, are slippery or insubstantial or somehow inaccessible to us. Outside the realm of what we are sure of lies the puzzling region of probable knowledge – puzzling in part because the sense of the noun seems to be cancelled by that of the adjective. The obvious move is to deny that the notion of knowledge has the importance generally attributed to it, and to try to make the concept of belief do the work that philosophers have generally assigned the grander concept. I shall argue that this is the right move.

A Pragmatic Analysis of Belief

To begin, we must get clear about the relevant sense of 'belief'. Here I follow Ramsey: 'the kind of measurement of belief with which probability is concerned is ... a measurement of belief *qua* basis of action'.[1]

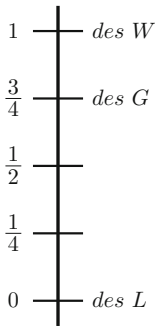
[1]Frank P. Ramsey, 'Truth and probability', in *The Foundations of Mathematics and Other Logical Essays*, R. B. Braithwaite, ed., London and New York, 1931, p. 171.

R.C. Jeffrey (deceased)
Princeton University, Boston, MA, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_4

Fig. 4.1



Ramsey's basic idea was that the desirability of a gamble G is a weighted average of the desirabilities of winning and of losing in which the weights are the probabilities of winning and of losing. If the proposition gambled upon is A, if the prize for winning is the truth of a proposition W, and if the penalty for losing is the truth of a proposition L, we then have

$$prob\ A = \frac{des\ G - des\ L}{des\ W - des\ L}. \tag{4.1}$$

Thus, if the desirabilities of losing and of winning happen to be 0 and 1, we have $prob\ A = des\ G$, as illustrated in Fig. 4.1, for the case in which the probability of winning is thought to be $\frac{3}{4}$.

On this basis, Ramsey[2] is able to give rules for deriving the gambler's subjective probability *and* desirability functions from his preference ranking of gambles, provided the preference ranking satisfies certain conditions of consistency. The probability function obtained in this way is a probability measure in the technical sense that, given any finite set of pairwise incompatible propositions which together exhaust all possibilities, their probabilities are non-negative real numbers that add up to 1. And in an obvious sense, probability so construed is a measure of the subject's willingness to act on his beliefs in propositions: it is a measure of degree of belief.

I propose to use what I take to be an improvement of Ramsey's scheme, in which the work that Ramsey does with the operation of forming gambles is done with the usual truth-functional operations on propositions.[3] The basic move is to restrict attention to certain 'natural' gambles, in which the prize for winning is the truth of the proposition gambled upon, and the penalty for losing is the falsity of that proposition. In general, the situation in which the gambler takes himself to be gambling on A with prize W and loss L is one in which he believes the proposition

[2] 'Truth and probability', F. P. Ramsey, *op. cit.*

[3] See Richard C. Jeffrey, *The Logic of Decision*, McGraw-Hill, 1965, the mathematical basis for which can be found in Ethan Bolker, *Functions Resembling Quotients of Measures*, Ph. D. Dissertation, Harvard University, 1965, and *Trans. Am. Math. Soc*., 124, 1966, pp. 293–312.

$$G = AW \vee \overline{A}L.$$

If G is a natural gamble we have $W = A$ and $L = \overline{A}$, so that G is the necessary proposition, $T = A \vee \overline{A}$:

$$G = A\,A \vee \overline{A}\,\overline{A} = T.$$

Now if A is a proposition which the subject thinks good (or bad) in the sense that he places it above T (or below T) in his preference ranking, we have

$$prob\ A = \frac{des\ T - des\ \overline{A}}{des\ A - des\ \overline{A}}, \tag{4.2}$$

corresponding to Ramsey's formula (4.1).

Here the basic idea is that if $A_1, A_2, \ldots, A_n$ are an exhaustive set of incompatible ways in which the proposition A can come true, the desirability of A must be a weighted average of the desirabilities of the ways in which it can come true:

$$des\ A = w_1\ des\ A_1 + w_2\ des\ A_2 + \cdots + w_n\ des\ A_n, \tag{4.3}$$

where the weights are the conditional probabilities,

$$w_i = prob\quad A_i/prob\quad A. \tag{4.4}$$

Let us call a function *des* which attributes real numbers to propositions a *Bayesian desirability function* if there is a probability measure *prob* relative to which (4.3) holds for all suitable $A, A_1, A_2, \ldots, A_n$. And let us call a preference ranking of propositions *coherent* if there is a Bayesian desirability function which ranks those propositions in order of magnitude exactly as they are ranked in order of preference. One can show[4] that if certain weak conditions are met by a coherent preference ranking, the underlying desirability function is determined up to a fractional linear transformation, i.e., if *des* and *DES* both rank propositions in order of magnitude exactly as they are ranked in order of preference, there must be real numbers a, b, c, d such that for any proposition A in the ranking we have

$$DES\ \ A = \frac{a\ des\ A + b}{c\ des\ A + d}. \tag{4.5}$$

The probability measure *prob* is then determined by (4.2) up to a certain quantization. In particular, if *des* is Bayesian relative to *prob*, then *DES* will be Bayesian relative to *PROB*, where

[4] Jeffrey, *op. cit.*, chs. 6, 8.

$$PROB\, A = prob\, A\ \ (c\ des\, A + d)\,. \tag{4.6}$$

Under further plausible conditions, (4.5) and (4.6) are given either exactly (as in Ramsey's theory) or approximately by

$$DES\, A = a\ des\, A + b, \tag{4.7}$$

$$PROB\, A = prob\, A. \tag{4.8}$$

I take the principal advantage of the present theory over Ramsey's to be that here we work with the subject's actual beliefs, whereas Ramsey needs to know what the subject's preference ranking of relevant propositions would be if his views of what the world is were to be changed by virtue of his having come to believe that various arbitrary and sometimes bizarre causal relationships had been established via gambles.[5]

To see more directly how preferences may reflect beliefs in the present system, observe that by (4.2) we must have $prob\, A > prob\, B$ if the relevant portion of the preference ranking is

$$\begin{array}{cc} A, & B \\ & T \\ & \overline{B} \\ & \overline{A} \end{array}$$

In particular, suppose that A and B are the propositions that the subject will get job 1 and that he will get job 2, respectively. Pay, working conditions, etc., are the same, so that he ranks A and B together. Now if he thinks himself more likely to get job 1 than job 2, he will prefer a guarantee of $(\overline{B})$ not getting job 2 to a guarantee of $(\overline{A})$ not getting job 1; for he thinks that an assurance of not getting job 2 leaves him more likely to get one or the other of the equally liked jobs than would an assurance of not getting job 1.

Probabilistic Acts and Observations

We might call a proposition *observational* for a certain person at a certain time if at that time he can make an observation of which the *direct* effect will be that his degree of belief in the proposition will change to 0 or to 1. Similarly, we might call a proposition *actual* for a certain person at a certain time if at that time he can perform an act of which the *direct* effect will be that his degree of belief in the proposition

[5] Jeffrey, *op. cit.,* pp. 145–150.

will change to 0 or to 1. Under ordinary circumstances, the proposition that the sun is shining is observational and the proposition that the agent blows his nose is actual. Performance of an act may give the agent what Anscombe calls[6] 'knowledge without observation' of the truth of an appropriate actual proposition. Apparently, a proposition can be actual or observational without the agent's knowing that it is; and the agent can be mistaken in thinking a proposition actual or observational.

The point and meaning of the requirement that the effect be 'direct', in the definitions of 'actual' and 'observational', can be illustrated by considering the case of a sleeper who awakens and sees that the sun is shining. Then one might take the observation to have shown him, directly that the sun is shining, and to have shown him indirectly that it is daytime. In general, an observation will cause numerous changes in the observer's belief function, but many of these can be construed as consequences of others. If there is a proposition E such that the *direct* effect of the observation is to change the observer's degree of belief in E to 1, then for any proposition A in the observer's preference ranking, his degree of belief in A after the observation will be the conditional probability

$$prob_E\, A = prob\ (A/E) = prob\, AE/prob\, E, \tag{4.9}$$

where *prob* is the observer's belief function before the observation. And conversely, if the observer's belief function after the observation is $prob_E$ and $prob_E$ is not identical with *prob*, then the *direct* effect of the observation will be to change the observer's degree of belief in E to 1. This completes a definition of *direct*.

But from a certain strict point of view, it is rarely or never that there is a proposition for which the direct effect of an observation is to change the observer's degree of belief in that proposition to 1; and from that point of view, the classes of propositions that count as observational or actual in the senses defined above are either empty or as good as empty for practical purposes. For if we care seriously to distinguish between 0.999 999 and 1.000 000 as degrees of belief, we may find that, after looking out the window, the observer's degree of belief in the proposition that the sun is shining is not quite 1, perhaps because he thinks there is one chance in a million that he is deluded or deceived in some way; and similarly for acts where we can generally take ourselves to be at best *trying* (perhaps with very high probability of success) to make a certain proposition true.

One way in which philosophers have tried to resolve this difficulty is to postulate a phenomenalistic language in which an appropriate proposition E can always be expressed, as a report on the immediate content of experience; but for excellent reasons, this move is now in low repute.[7] The crucial point is not that 0.999 999 is so close to 1.000 000 as to make no odds, practically speaking, for situations abound in which the gap is more like one half than one millionth. Thus, in examining a piece of cloth by candlelight one might come to attribute probabilities 0.6 and 0.4

[6] G. E. M. Anscombe, *Intention,* § 8, Oxford, 1957; 2nd ed., Ithaca, N.Y., 1963.

[7] See, e.g., J. L. Austin, *Sense and Sensibilia*, Oxford, 1962.

to the propositions G that the cloth is green and B that it is blue, without there being any proposition E for which the direct effect of the observation is anything near changing the observer's degree of belief in E to 1. One might think of some such proposition as that (E) *the cloth looks green or possibly blue*, but this is far too vague to yield $prob\left(\frac{G}{E}\right) = 0.6$ and $prob\left(\frac{B}{E}\right) = 0.4$. Certainly, there is *something* about what the observer sees that leads him to have the indicated degrees of belief in G and in B, but there is no reason to think the observer can express this something by a statement in his language. And physicalistically, there is some perfectly definite pattern of stimulation of the rods and cones of the observer's retina which prompts his belief, but there is no reason to expect him to be able to describe that pattern or to recognize a true description of it, should it be suggested.

As Austin[8] points out, the crucial mistake is to speak seriously of the *evidence* of the senses. Indeed the relevant experiences have perfectly definite characteristics by virtue of which the observer comes to believe as he does, and by virtue of which in our example he comes to have degree of belief 0.6 in G. But it does not follow that there is a proposition E of which the observer is certain after the observation and for which we have $prob\left(\frac{G}{E}\right) = 0.6$, $prob\left(\frac{B}{E}\right) = 0.4$, etc.

In part, the quest for such phenomenological certainty seems to have been prompted by an inability to see how uncertain evidence can be used. Thus C. I. Lewis:

> If anything is to be probable, then something must be certain. The data which themselves support a genuine probability, must themselves be certainties. We do have such absolute certainties, in the sense data initiating belief and in those passages of experience which later may confirm it. But neither such initial data nor such later verifying passages of experience can be phrased in the language of objective statement – because what can be so phrased is never more than probable. Our sense certainties can only be formulated by the expressive use of language, in which what is signified is a content of experience and what is asserted is the givenness of this content.[9]

But this motive for the quest is easily disposed of.[10] Thus, in the example of observation by candlelight, we may take the direct result of the observation (in a modified sense of 'direct') to be, that the observer's degrees of belief in G and B change to 0.6 and 0.4. Then his degree of belief in any proposition A in his preference ranking will change from $prob\ A$ to

$$PROB\ A = 0.6\ prob\,(A/G) + 0.4\ prob\,(A/B)\,.$$

In general, suppose that there are propositions $E_1, E_2, \ldots, E_n$, in which the observer's degrees of belief after the observation are $p_1, p_2, \ldots, p_n$; where the E's are pairwise incompatible and collectively exhaustive; where for each i, $prob\ E_i$ is

[8] Austin, *op. cit.,* ch. 10

[9] C. I. Lewis, *An Analysis of Knowledge and Valuation,* La Salle, Illinois, 1946, p. 186.

[10] Jeffrey, *op. cit.,* ch. 11.

neither 0 nor 1; and where for each proposition A in the preference ranking and for each i the conditional probability of A on E_i is unaffected by the observation:

$$PROB\ (A/E_i) = prob\ (A/E_i)\,. \tag{4.10}$$

Then the belief function after the observation may be taken to be *PROB*, where

$$PROB\,A = p_1\ prob\ (A/E_1) + p_2\ prob\ (A/E_2) + \cdots + p_n\ prob\,(A/E_n)\,, \tag{4.11}$$

if the observer's preference rankings before and after the observation are both coherent. Where these conditions are met, the propositions $E_1, E_2, \ldots, E_n$, may be said to form a *basis* for the observation; and the notion of a basis will play the role vacated by the notion of *directness*.

The situation is similar in the case of acts. A marksman may have a fairly definite idea of his chances of hitting a distant target, e.g. he may have degree of belief 0.3 in the proposition H that he will hit it. The basis for this belief may be his impressions of wind conditions, quality of the rifle, etc.; but there need be no reason to suppose that the marksman can express the relevant data; nor need there be any proposition E in his preference ranking in which the marksman's degree of belief changes to 1 upon deciding to fire at the target, and for which we have $prob\left(\frac{H}{E}\right) = 0.3$. But the pair $H, \overline{H}$ may constitute a *basis* for the act, in the sense that for any proposition A in the marksman's preference ranking, his degree of belief after his decision is

$$PROB\,A = 0.3\ prob\ (A/H) + 0.7\ prob\ \left(A/\overline{H}\right).$$

It is correct to describe the marksman as *trying* to hit the target; but the proposition that he is trying to hit the target can not play the role of E above. Similarly, it was correct to describe the cloth as *looking* green or possibly blue; but the proposition that the cloth looks green or possibly blue does not satisfy the conditions for directness.

The notion of directness is useful as well for the resolution of unphilosophical posers about probabilities, in which the puzzling element sometimes consists in failure to think of an appropriate proposition E such that the direct effect of an observation is to change degree of belief in E to 1, e.g. in the following problem reported by Mosteller.[11]

> Three prisoners, *a, b*, and *c*, with apparently equally good records have applied for parole. The parole board has decided to release two of the three, and the prisoners know this but not which two. A warder friend of prisoner *a* knows who are to be released. Prisoner *a* realizes that it would be unethical to ask the warder if he, *a*, is to be released, but thinks of asking for the name of *one* prisoner *other than himself* who is to be released. He thinks that before he asks, his chances of release are $\frac{2}{3}$. He thinks that if the warder says '*b* will be released,' his own chances have now gone down to $\frac{1}{2}$, because either *a* and *b* or *b* and *c* are to be released. And so *a* decides not to reduce his chances by asking. However, *a* is mistaken in his calculations. Explain.

[11] Problem 13 of Frederick Mosteller, *Fifty Challenging Problems in Probability,* Reading, Mass., Palo Alto, and London, 1965.

Here indeed the possible cases (in a self-explanatory notation) are

$$AB,\ AC,\ BC,$$

and these are viewed by a as equiprobable. Then *prob* A is $\frac{2}{3}$ but *prob* $\left(\frac{A}{B}\right) = prob\left(\frac{A}{B}\right) = \frac{1}{2}$, and, since the warder must answer either 'b' or 'c' to a's question, it looks as if the direct result of the 'observation' will be that a comes to attribute probability 1 either to the proposition B that b will be released, or to the proposition C that c will be released. But this is incorrect. The relevant evidence-proposition would be more like the proposition *that the warder says,* 'b', or *that the warder says*, 'c', even though neither of these will quite do. For it is only in cases AB and AC that the warder's reply is dictated by the facts: in case BC, where b and c are both to be released, the warder must somehow choose *one* of the two true answers. If a expects the warder to make the choice by some such random device as tossing a coin, then we have *prob* (A/the warder says, 'b') = *prob* (A/the warder says, 'c') = *prob* $A = \frac{2}{3}$; while if a is sure that the warder will say 'b' if he can, we have *prob* (A/the warder says 'b') $= \frac{1}{2}$ but *prob* (A/the warder says 'c') = 1.

3. *Belief*: *reasons vs. causes*. Indeed it is desirable, where possible, to incorporate the results of observation into the structure of one's beliefs via a basis of form $E, \bar{E}$ where the probability of E after the observation is nearly 1. For practical purposes, E then satisfies the conditions of directness, and the 'direct' effect of the observation can be described as informing the observer of the truth of E. Where this is possible, the relevant passage of sense experience *causes* the observer to believe E; and if *prob* $\left(\frac{A}{E}\right)$ is high, his belief in E may be a *reason* for his believing A, and E may be spoken of as (inconclusive) *evidence* for A. But the sense experience is evidence neither for E nor for A. Nor does the situation change when we speak physicalistically in terms of patterns of irritation of our sensory surfaces, instead of in terms of sense experience: such patterns of irritation *cause* us to believe various propositions to various degrees; and sometimes the situation can be helpfully analyzed into one in which we are caused to believe $E_1, E_2, \ldots, E_n$, to degrees $p_1, p_2, \ldots, p_n$, whereupon those beliefs provide *reasons* for believing other propositions to other degrees. But patterns of irritation of our sensory surfaces are not reasons or evidence for any of our beliefs, any more than irritation of the mucous membrane of the nose is a *reason* for sneezing.

When I stand blinking in bright sunlight, I can no more believe that the hour is midnight than I can fly. My degree of belief in the proposition that the sun is shining has two distinct characteristics, (a) It is 1, as close as makes no odds. (b) It is compulsory. Here I want to emphasize the second characteristic, which is most often found in conjunction with the first, but not always. Thus, if I examine a normal coin at great length, and experiment with it at length, my degree of belief in the proposition that the next toss will yield a head will have two characteristics, (a) It

is $\frac{1}{2}$. (b) It is compulsory. In the case of the coin as in the case of the sun, I cannot decide to have a different degree of belief in the proposition, any more than I can decide to walk on air.

In my scientific and practical undertakings I must make use of such compulsory beliefs. In attempting to understand or to affect the world, I cannot escape the fact that I am part of it: I must rather make use of that fact as best I can. Now where epistemologists have spoken of observation as a source of *knowledge*, I want to speak of observation as a source of compulsory *belief* to one or another degree. I do not propose to identify a very high degree of belief with knowledge, any more than I propose to identify the property of being near 1 with the property of being compulsory.

Nor do I postulate any *general* positive or negative connection between the characteristic of being compulsory and the characteristic of being sound or appropriate in the light of the believer's experience. Nor, finally, do I take a compulsory belief to be necessarily a permanent one: new experience or new reflection (perhaps, prompted by the arguments of others) may loosen the bonds of compulsion, and may then establish new bonds; and the effect may be that the new state of belief is sounder than the old. or less sound.

Then why should we trust our beliefs? According to K. R. Popper,

> . . . the decision to accept a basic statement, and to be satisfied with it, is causally connected with our experiences – especially with our *perceptual experiences*. But we do not attempt to *justify* basic statements by these experiences. Experiences can *motivate a decision*, and hence an acceptance or a rejection of a statement, but a basic statement cannot be *justified* by them – no more than by thumping the table.[12]

I take this objection to be defective, principally in attempting to deal with basic statements (observation reports) in terms of *decisions* to *accept* or to *reject* them. Here acceptance parallels belief, rejection parallels disbelief (belief in the denial), and tentativeness or reversibility of the decision parallels *degree* of belief. Because logical relations hold between statements, but not between events and statements, the relationship between a perceptual experience (an event of a certain sort) and a basic statement cannot be a logical one, and therefore, Popper believes, cannot be of a sort that would justify the statement:

> Basic statements are accepted as the result of a decision or agreement; and to that extent they are conventions.[13]

But in the absence of a positive account of the nature of acceptance and rejection, parallel to the account of partial belief given in section 1, it is impossible to evaluate this view. Acceptance and rejection are apparently acts undertaken as results of decisions; but somehow the decisions are conventional – perhaps only in the sense that they may be *motivated* by experience, but not *adequately* motivated, if adequacy entails justification.

---

[12] K. R. Popper, *The Logic of Scientific Discovery,* London, 1959, p. 105.

[13] Popper, *op. cit.,* p. 106.

To return to the question, 'Why should we trust our beliefs?' one must ask what would be involved in *not* trusting one's beliefs, if belief is analyzed as in section 1 in terms of one's preference structure. One way of mistrusting a belief is declining to act on it, but this appears to consist merely in lowering the degree of that belief: to mistrust a partial belief is then to alter its degree to a new, more suitable value.

A more hopeful analysis of such mistrust might introduce the notion of sensitivity to further evidence or experience. Thus, agents 1 and 2 might have the same degree of belief – $\frac{1}{2}$ – in the proposition H_1 that the first toss of a certain coin will yield a head, but agent 1 might have this degree of belief because he is convinced that the coin is normal, while agent 2 is convinced that it is either two-headed or two-tailed, he knows not which.[14] There is no question here of agent 2's expressing his mistrust of the figure $\frac{1}{2}$ by lowering or raising it, but he can express that mistrust quite handily by aspects of his belief function. Thus, if H_i is the proposition that the coin lands head up the ith time it is tossed, agent 2's beliefs about the coin are accurately expressed by the function $prob_2$ where

$$prob_2\ H_i = \frac{1}{2},\ prob_2\ \left(H_i/H_j\right) = 1,$$

while agent 1's beliefs are equally accurately expressed by the function $prob_1$ where

$$prob_1\ \left(H_{i_1},\ H_{i_2}, \ldots, H_{i_n}\right) = 2^{-n},$$

if $i_1 < i_2 < \ldots < i_n$. In an obvious sense, agent 1's beliefs are *firm* in the sense that he will not change them in the light of further evidence, since we have

$$prob_1\ (H_{n+1}/H_1,\ H_2, \ldots,\ H_n) = prob_1\ H_{n+1} = \frac{1}{2},$$

while agent 2's beliefs are quite tentative and in that sense, mistrusted by their holder. Still, $prob_1\ H_i = prob_2\ H_i = \frac{1}{2}$.

After these defensive remarks, let me say how and why I take compulsive belief to be sound, under appropriate circumstances. Bemused with syntax, the early logical positivists were chary of the notion of truth; and then, bemused with Tarski's account of truth, analytic philosophers neglected to inquire how we come to believe or disbelieve simple propositions. Quite simply put, the point is: coming to have suitable degrees of belief in response to experience is a matter of training – a *skill* which we begin acquiring in early childhood, and are never quite done polishing. The skill consists not only in coming to have appropriate degrees of belief in appropriate propositions under paradigmatically good conditions of observation, but also in coming to have appropriate degrees of belief between zero and one when conditions are less than ideal.

Thus, in learning to use English color words correctly, a child not only learns to acquire degree of belief 1 in the proposition that the cloth is blue, when in bright sunlight he observes a piece of cloth of uniform hue, the hue being squarely in

[14] This is a simplified version of 'the paradox of ideal evidence', Popper, *op. cit.,* pp. 407–409.

the middle of the blue interval of the color spectrum: he also learns to acquire appropriate degrees of belief between 0 and 1 in response to observation under bad lighting conditions, and when the hue is near one or the other end of the blue region. Furthermore, his understanding of the English color words will not be complete until he understands, in effect, that blue is between green and violet in the color spectrum: his understanding of this point or his lack of it will be evinced in the sorts of mistakes he does and does not make, e.g. in mistaking green for violet he may be evincing confusion between the meanings of 'blue' and of 'violet', in the sense that his mistake is linguistic, not perceptual.

Clearly, the borderline between factual and linguistic error becomes cloudy, here: but cloudy in a perfectly realistic way, corresponding to the intimate connection between the ways in which we experience the world and the ways in which we speak. It is for this sort of reason that having the right language can be as important as (and can be in part identical with) having the right theory.

Then learning to use a language properly is in large part like learning such skills as riding bicycles and flying aeroplanes. One must train oneself to have the right sorts of responses to various sorts of experiences, where the responses are degrees of belief in propositions. This may, but need not, show itself in willingness to utter or assent to corresponding sentences. Need not, because e.g. my cat is quite capable of showing that it thinks it is about to be fed, just as it is capable of showing what its preference ranking is, for hamburger, tuna fish, and oat meal, without saying or understanding a word. With people as with cats, evidence for belief and preference is behavioral; and speech is far from exhausting behavior.[15]

Our degrees of beliefs in various propositions are determined jointly by our training and our experience, in complicated ways that I cannot hope to describe. And similarly for conditional subjective probabilities, which are certain ratios of degrees of belief: to some extent, these are what they are because of our training – because we speak the languages we speak. And to this extent, conditional subjective probabilities reflect *meanings*. And in this sense, there can be a theory of degree of confirmation which is based on analysis of meanings of sentences. Confirmation theory is therefore semantical and, if you like, logical.[16]

Discussion

L. HURWICZ: *Richard Jeffrey on the Three Prisoners*.

I would like to make a comment which I think is along the lines of Professor Jeffrey's discussion of the three prisoners. I would like to make the situation a little

[15] Jeffrey, *op. cit*., pp. 57–59.

[16] Support of U.S. Air Force Office of Scientific Research is acknowledged, under Grant AF–AFOSR–529–65.

more explicit than it was earlier, although I shall not contradict anything that has been said: I think this will help us to see to what extent, if any, there is anything surprising or paradoxical in the situation.

First of all let me say this: there were three possible decisions by the warden – *AB, BC* and *AC*; then, as against that, there was also the question of what the warden would say to *a* who asked the question who else was being freed, and clearly the warden could only answer '*b*' or '*c*'. What I'm going to put down here is simply the bivariate or two-way probability distribution, and it doesn't matter at all at this stage whether we interpret it as a frequency or as a subjective probability, because it's just a matter of applying the mechanics of the Bayes theorem.

One other remark I'd like to make is this: the case that was considered by Professor Jeffrey was one where the *a priori* probabilities of *AB, BC* and *AC* were each one-third. This actually does not at all affect the reasoning, and I will stick with it just because it is close to my limitations in arithmetic.

So the marginal frequencies or probabilities are all equal to one-third. If the decision had been *AB*, then of course the warden could only answer '*b*', and similarly if the decision had been *AC*, he could only answer '*c*'. So the joint frequency or probability of the following event is one-third: the people chosen for freeing are *a* and *b*, and when the warden is asked, 'Who is the person other than *a* who is about to be freed?', his answer is '*b*'. The joint probability is also one-third that the choice was *AC* and that the warden answered '*c*'.

We now come to the only case where the warden has a choice of what he will say, namely, the case where the decision was *BC*. The question was raised, quite properly, of how he goes about making this choice.

Let me here say the following. In a sense what I'm doing here is a sally into enemy territory, because I personally am not particularly Bayesian in my approach to decision theory, so I would not myself assert that the only method is to describe the warden's decision, the warden's principle of choice, as a probabilistic one. However, if it is not probabilistic, then of course the prisoner, our *a*, would have to be using some other principle of choice on his part in order to decide what to do. Being an unrepentant conservative on this, I might choose, or *A* might choose, the minimax principle. However, in order to follow the spirit of the discussion here, I will assume that the whole thing is being done in a completely Bayesian or probabilistic way; in this case, to compute the remaining joint distribution we must make some probabilistic assumption about how the warden will behave when asked the question.

So let the principle be this, that he has a certain random device such that if the people to be freed are *b* and *c*, his answer to the question will be '*b*' with probability β and '*c*' with of course probability $1 - \beta$. All I will assume for the moment about β is that it is between zero and one, and that's probably one of the few uncontroversial points so far.

It is clear that the sum of the two joint probabilities (*BC* and '*b*', and *BC* and '*c*') will be one-third; so the first will be $\frac{1}{3}\beta$, and the second $\frac{1}{3}(1 - \beta)$. The marginal (or absolute) probabilities of '*b*' and '*c*' will be $\frac{1}{3}(1 + \beta)$ and $\frac{1}{3}(2 - \beta)$ respectively.

Inf. → Dec. ↓	'b'	'c'	Marginal
AB	$\frac{1}{3}$	0	$\frac{1}{3}$
BC	$\beta/3$	$(1-\beta)/3$	$\frac{1}{3}$
AC	0	$\frac{1}{3}$	$\frac{1}{3}$
Marginal	$(1+\beta)/3$	$(2-\beta)/3$	1

Now what are the probabilities after the warden has given his answer? Suppose that the answer that the warden gave is 'b': the problem now is, what is the probability that a is to be freed, given that the warden said that b is to be freed? This probability, which I will denote by 'π_b', I obtain in the following way using what I hope is a self-explanatory notation:

$$\begin{aligned}
\pi_b &= p\,(A/\text{`}b\text{'}) \\
&= \frac{p\,(A\cdot\text{`}b\text{'})}{p\,(\text{`}b\text{'})} \\
&= \frac{p\,(AB\cdot\text{`}b\text{'}) + p\,(AC\cdot\text{`}b\text{'})}{p\,(AB\cdot\text{`}b\text{'}) + p\,(AC\cdot\text{`}b\text{'}) + p\,(BC\cdot\text{`}b\text{'})} \\
&= \frac{\frac{1}{3}+0}{\frac{1}{3}+0+\beta/3} \\
&= 1/\,(1+\beta)\,.
\end{aligned}$$

Similarly I get π_c (the probability that a is to be freed, given that the warden said that b is to be freed) as follows:

$$\begin{aligned}
\pi_c &= p\,(A/\text{`}c\text{'}) \\
&= \frac{p\,(A\cdot\text{`}c\text{'})}{p\,(\text{`}c\text{'})} \\
&= \frac{p\,(AB\cdot\text{`}c\text{'}) + p\,(AC\cdot\text{`}c\text{'})}{p\,(AB\cdot\text{`}c\text{'}) + p\,(AC\cdot\text{`}c\text{'}) + p\,(BC\cdot\text{`}c\text{'})} \\
&= \frac{0+\frac{1}{3}}{0+\frac{1}{3}+(1-\beta)\,/3} \\
&= 1/\,(2-\beta)\,.
\end{aligned}$$

Now the question which we now have to ask is this: are these conditional probabilities, π, different from the marginal (absolute) probability that a is to be freed, $p(a)$?

And the answer is that they are except when β happens to be equal to one-half, in which case the probability remains at its marginal value of two-thirds. But except in this special case the probabilities π_b and π_c can vary from one-half to one.[17]

As I indicated before, there is no quarrel between us, but I do want to explore just one step further, and that is this. You remember when we were told this anecdote there was a wave of laughter and I now want to see what it was that was so funny. It is that this prisoner became doubtful about asking for this extra information, because he thought his probability of being released would go down after getting it. So it seemed that having this extra information would make him less happy, even though he didn't have to pay for it. That really was the paradox, not the fact that the probabilities changed. Clearly, the change in probabilities is itself not at all surprising; for example, if the warden had told *a* the names of *two* people other than himself who would be freed, his optimism would have gone down very drastically.[18]

What is surprising is that *a* thought he would be less happy with the *prospect* of having the extra piece of information than without this prospect. What I want to show now is that *a* was just wrong to think this; in other words, if this information was free, he should have been prepared to hear it.

Suppose for instance β is different from one-half: I think it is implicit in this little anecdote that the probability of *a*'s being released either before or after getting the information, in some sense corresponds to his level of satisfaction. If his chances are good he is happy; if his chances are bad he is miserable. So these π's, though they happen to have been obtained as probabilities, may at the same time be interpreted as utilities or what Professor Jeffrey called desirabilities. Good. Now if *a* proceeds in the Bayesian way he has to do the following: he has to look at all these numbers, because before he asks for the information he does not know whether the answer will be '*b*' or '*c*'. Then he must ask himself the following: How happy will I be if he says '*b*'? How happy will I be if he says '*c*'? And then in the Bayesian (or de Finetti or Ramsey) spirit he multiplies the utilities, say $u(\text{'}b\text{'})$ and $u(\text{'}c\text{'})$ associated with hearing the warden say '*b*' or '*c*' by the respective probabilities, say $p(\text{'}b\text{'})$ and $p(\text{'}c\text{'})$, of hearing these answers. He thus obtains an expression for his expected[19] utility associated with getting the extra information, say

$$Eu^* = p\left('b'\right) \cdot u\left('b'\right) + p\left('c'\right) \cdot u\left('c'\right).$$

[17] In the problem as reported by Mosteller, it might be reasonable to take $\beta = \frac{1}{2}$. In that case, let us note, $\pi_b = 1/\left(1 + \frac{1}{2}\right) = \frac{2}{3}$ (not $\frac{1}{2}$ as suggested in the statement of the problem!) and also $\pi_c = 1/(2 - \beta) = 1/\left(2 - \frac{1}{2}\right) = \frac{2}{3}$. Hence (for $\beta = \frac{1}{2}$) *a* was wrong to expect the probabilities to change. But, on the other hand, the warden's reply would give him no additional information.

[18] Or suppose, that $\beta = 1$ (and *a* knows this). Then if *a* hears the warden tell him that *c* is one of the persons to be released, he will have good reason to feel happy. For when $\beta = 1$, the warden will tell *a* about having selected *c* only if the selected pair was *AC*. On the other hand, still with $\beta = 1$, if the warden says that *b* is one of the persons to be released, this means (with equal probabilities) that either *AB* or *BC* has been chosen, but *not AC*. Hence, with the latter piece of information, *a* will be justifiably less optimistic about his chances of release. (With β close to one, a similar situation prevails.)

[19] In the sense of the mathematical expectation of a random variable.

Now the required probabilities are the marginal probabilities at the bottom of the table, i.e.,

$$p\left('b'\right) = \frac{1+\beta}{3},\ p\left('c'\right) = \frac{2-\beta}{3}.$$

As for utilities, it is implicit in the argument that they are linear[20] functions of the probabilities that a will be released given the warden's answer. So

$$u\left('b'\right) = \pi_b = \frac{1}{1+\beta},\ u\left('c'\right) = \pi_c = \frac{1}{2-\beta}.$$

Hence the (expected) utility associated with getting the extra information from the warden is

$$Eu^* = \frac{1+\beta}{3}.\frac{1}{1+\beta} + \frac{2-\beta}{3}.\frac{1}{2-\beta} = \frac{2}{3}.$$

On the other hand, the expected utility $Eu°$, associated with *not* asking the warden for extra information is simply equal to the original probability $p(a)$ that a will be released,

$$Eu° = p(a) = \frac{2}{3}.$$

Hence it so happens that (for a utility function linear in probabilities of release)

$$Eu^* = Eu°,$$

i.e., the expected utility with extra information (Eu^*) is the same as without extra information ($Eu°$). Thus a should be willing (but not eager) to ask for extra information (if it is free of charge). 'On the average'[21], it won't do him any harm; nor will it help him.[22]

P. SUPPES: *Rational Changes of Belief.*

I am generally very much in agreement with Professor Jeffrey's viewpoint on belief and knowledge as expressed in his paper. The focus of my brief remarks is to point out how central and difficult are the problems concerning *changes* of belief. Jeffrey remarks that the familiar method of changing probable beliefs by explicitly conditionalizing the relevant probability measure is not adequate in many

[20] See footnote 22 on the next page.

[21] 'On the average' expresses the fact that the decision is made on the basis of mathematical expectation. It need not imply a frequency interpretation of probabilities.

[22] When utilities are *non-linear* with respect to probabilities of release, the prospect of additional information may be helpful or harmful.

situations – in fact, in all those situations that involve a change in the probability assigned to evidence, but a change that does not make the probability of possible evidence 0 or 1.

My point is that once we acknowledge this fact about the probable character of evidence we open Pandora's box of puzzles for any theory of rational behavior. I would like to mention three problems. These problems are not dealt with explicitly by Jeffrey, but the focus I place on them is certainly consistent with his own expressed views.

Attention and Selection

I begin with the problem of characterizing how a man attempting to behave rationally is to go about selecting and attending to what seems to be the right kind of evidence. Formulations of the problem of evidence within a logically well-defined language or sample space have already passed over the problem of evaluating their appropriateness. Any man has highly limited capacities for attention and observation. Given a characterization of his powers and limitations what is a rational way to behave? Consider the familiar coin-flipping example. When is it appropriate to stop concentrating on the outcomes of the flips and to start observing closely the behavior of the gambler doing the flipping? To take another sort of example, how do we characterize the behavior of detectives who act on subthreshold cues and conjectures that can scarcely even be verbalized, let alone supported by explicit evidence? Put more generally, what is the rational way to go about discovering facts as opposed to verifying them? It is easy to claim that for a wide variety of situations rational discovery is considerably more important than rational verification. In these cases of discovery we need to understand much better the information-processing capacities of human beings in order to be able to characterize in a serious way *rational* information-processing. Above all, it is certainly not clear to me what proportion of rational information-processing should be verbalized or is even verbalizable.

Finite Memory

The second problem concerns the rational use of our inevitably restricted memory capacities. A full-blown theory of rationality should furnish guidelines for the kinds of things that should be remembered and the kind that should not. Again a solution, but certainly not a solution whose optimality can be seriously defended, is at least partly given by the choice of a language or a sample space for dealing with a given set of phenomena. But the amount of information impinging on any man in a day or a week or a month is phenomenal and what is accessible if he chooses to make it so is even more so. What tiny fraction of this vast potential should be absorbed and

stored for ready access and use? Within the highly limited context of mathematical statistics, certain answers have been proposed. For example, information about the outcome of an experiment can be stored efficiently and with little loss of information in the form of the likelihood function or some other sufficient statistic, but this approach is not of much use in most situations, although elements of the approach can perhaps be generalized to less restricted circumstances. Perhaps even more importantly, it is not clear what logical structure is the most rational to impose on memory. The attempts at constructing associative computer memories, as they are often called, show how little we are able as yet to characterize explicitly a memory with the power and flexibility of a normal man's, not to speak of the memory of a normal man who is using his powers with utmost efficiency. Perhaps one of the most important weaknesses of confirmation theory and the Ramsey-sort of theory developed by Jeffrey and others is that little is said about imposing a logical structure on evidence. Part of the reason for this is that the treatment of evidence is fundamentally static rather than dynamic and temporal. In real life, evidence accumulates over time and we tend to pay more attention to later than earlier data, but the appropriate logical mechanisms for storing, organizing and compressing temporally ordered data are as yet far from being understood.

Concept Formation

The most fundamental and the most far-reaching cognitive changes in belief undoubtedly take place when a new concept is introduced. The history of science and technology is replete with examples ranging from the wheel to the computer, and from arithmetic to quantum mechanics. Perhaps the deepest problem of rational behavior, at least from a cognitive or epistemological standpoint, is to characterize when a man should turn from using the concepts he has available to solve a given problem to the search not just for new evidence but for new concepts with which to analyze the evidence. Perhaps the best current example of the difficulty of characterizing the kinds of concepts we apply to the solution of problems is the floundering and vain searching as yet typical of the literature on artificial intelligence. We cannot program a computer to think conceptually because we do not understand how men think conceptually, and the problem seems too difficult to conceive of highly nonhuman approaches. For those of us interested in rational behavior the lesson to be learned from the tantalizing yet unsatisfactory literature on artificial intelligence is that we are a long way from being able to say what a rational set of concepts for dealing with a given body of experience should be like, for we do not have a clear idea of what conceptual apparatus we actually use in any real sense.

To the problems about rationality I have raised in these remarks there is the pat answer that these are not problems of the theory of rational behavior but only of the theory of actual behavior. This I deny. A theory of rationality that does not

take account of the specific human powers and limitations of attention, memory and conceptualization may have interesting things to say but not about human rationality.

R. C. JEFFREY: *Reply*.

Suppes' and Hurwicz' comments are interesting and instructive, and I find I have little to add to them. But perhaps a brief postscript is in order, in response to Suppes' closing remark:

> A theory of rationality that does not take account of the specific human powers and limitations of attention may have interesting things to say, but not about human rationality.

It may be that there is no real issue between us here, but the emphasis makes me uncomfortable. In my view, the logic of partial belief is a branch of decision theory, and I take decision theory to have the same sort of relevance to human rationality that (say) quantification theory has: the relevance is there, even though neither theory is directly about human rationality, and neither theory takes any account of the specific powers and limitations of human beings.

For definiteness, consider the following preference ranking of four sentences s, s', t, t', where s and s' are logically inconsistent, as are t and t'.

$$s$$
$$s'$$
$$t$$
$$t'$$

This ranking is *incoherent*: it violates at least one of the following two requirements, (a) Logically equivalent sentences are ranked together. (b) The disjunction of two logically incompatible sentences is ranked somewhere in the interval between them, endpoints included. Requirements (a) and (b) are part of (or anyway, implied by) a definition of 'incoherent'. To see that the given ranking is incoherent, notice that (a) implies that the disjunction of the sentences s, s' is ranked with the disjunction of the sentences t, t', while (b) implies that in the given ranking, the first disjunction is higher than the second. In my view, the point of classifying this ranking as incoherent is much like the point of classifying the pair s, s' as logically inconsistent: the two classifications have the same sort of relevance to human rationality. In the two cases, a rational man who made the classification would therefore decline to own the incoherent preference ranking or to believe both of the inconsistent sentences. (For simplicity I speak of belief here as an all-or-none affair.)

True enough: since there is no effective decision procedure for quantificational consistency there is no routine procedure a man can use – be he ever so rational – to correctly classify arbitrary rankings of sentences as incoherent or arbitrary sets of sentences as inconsistent. The relevance of incoherence and inconsistency to human rationality is rather that a rational man, once he comes to see that his preferences are incoherent or that his beliefs are inconsistent, will proceed to revise them. In carrying out the revision he may use decision theory or quantification theory as an aid; but neither theory fully determines how the revision shall go.

In fine, I take Bayesian decision theory to comprise a sort of *logic* of decision: the notion of coherence has much the same sort of relationship to human ideals of rationality that the notion of consistency has. But this is not to deny Suppes' point. The Bayesian theory is rather like a book of rules for chess which tells the reader what constitutes winning: there remains the question of ways and means.

Chapter 5
Fine-Grained Opinion, Probability, and the Logic of Full Belief

Bas C. van Fraassen

Personal or subjective probability entered epistemology as a cure for certain perceived inadequacies in the traditional notion of belief. But there are severe strains in the relationship between probability and belief. They seem too intimately related to exist as separate but equal; yet if either is taken as the more basic, the other may suffer.

After explaining the difficulties in some detail I will propose a single unified account which takes conditional personal probability as basic. Full belief is therefore a defined, derivative notion. Yet it is easiest to explain the resulting picture of opinion as follows: my subjective probability is only a grading of the possibilities left open by my beliefs. My conditional probabilities generally derive from the strongest belief I can maintain when admitting the relevant condition. Appendices will survey the literature.

Full Belief and Personal Probability

The most forceful answer I can give if asked for my opinion, is to say what I fully believe. The point of having beliefs is to construct a *single* (though in general incomplete) picture of what things are like. One obvious model of this part of my opinion is a set of propositions.[1] Their intersection is the proposition which captures

[1]This was essentially the model provided in Hintikka's book *Knowledge and Belief*. By propositions I mean the semantic content of statements; the same proposition can be expressed by many statements. I am not addressing how opinion is stored or communicated.

B.C. van Fraassen (✉)
San Francisco State University, San Francisco, CA, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_5

exactly that single picture of the world which has my full assent. Clearly a person's full beliefs leave open many alternatives. Alternatives left open by belief are then also represented by (sets of) propositions, namely ones that imply my beliefs. But these alternatives do not all have the same status for me, though they are all "possible for all I [know or] believe." Some seem more or less likely than others: enter personal (subjective) probability, as a grading of the possibilities left open by one's beliefs.

I will take for granted that the probability of a proposition is a real number in the interval [0, 1], with the empty proposition Λ (self-contradiction) receiving 0 and the universal proposition U (tautology) receiving 1. The assignment is a *measure*, that is, it is additive and continuous (equivalently, countably additive). It follows from this that the assignment of probabilities respects the ordering by logical implication:

$$\text{If } A \subseteq B \text{ then } P(A) \leq P(B)$$

though we must be careful in any extrapolation from propositions to sets of propositions unless they are countable. That is essentially because at most countably many disjoint propositions can receive finite positive probability. (*Reason*: at most one can receive probability greater than 1/2, at most two can receive more than 1/3, ... etc. The question of infinitesimal positive probability will be taken up in an Appendix.)

The so-called lottery paradox shows that we cannot equate belief with probability $\geq p$, if $p < 1$. For example, suppose $p = 0.99$ and a lottery which I believe to be fair has 1000 tickets, then my probability that the k^{th} ticket will not win the (single) prize equals 0.999. Hence for each $k = 1, \ldots, 1000$, I would believe that the k^{th} ticket will not win. My beliefs would then entail that all tickets will fail to win, which conflicts with my original belief that the lottery is fair. This argument is more important for what it presupposes than for what it shows. It is clearly based on the assumed role of full belief: to form a single, unequivocally endorsed picture of what things are like.[2]

In fact, the thesis that probabilities grade exactly the alternatives left open by full belief guarantees that all full beliefs have maximal personal probability.

So what if we simply set $p = 1$, that is, *identify* our full beliefs with the propositions which are maximally likely to us? The first problem with this is that we seem to be treating full beliefs as on a par with tautologies. Are there no distinctions to be drawn among the maximally likely propositions? There is a second problem for this proposal as well. In science we deal with continuous quantities. Therefore, in general, if we let science guide our opinion, the maximally likely propositions will not form a single picture – they will just give us a family of rival maximally likely pictures.

Example 5.1 Consider the mass of the moon reckoned in kilograms, and suppose I am sure that it is a number in the interval $[a, b]$. If my probability follows Lebesgue

[2]This has been denied, e.g. by Henry Kyburg, and doubted, e.g. by Richard Foley (1993, Ch. 4).

measure then my probability is *zero* that the number equals x, for $a \leq x \leq b$. Hence my probability equals 100 % that the number lies in the set $[a, b] - \{x\}$, for each such number x. Yet no real number belongs to all these sets – their intersection is empty. Probability measures of this sort (deriving from continuous probability densities) are ubiquitous in science, and informed opinion must be allowed to let itself be guided by them. We have here a transfinite lottery paradox, and we can't get out of it in the way that worked for the finite case (see Maher 1990).

Supposition and Two-Place Probability

There is a third aspect of opinion, besides belief and subjective grading, namely *supposition*. Much of our opinion can be elicited only by asking us to suppose something, which we may or may not believe. The respondent imaginatively puts himself in the position of someone for whom the supposition has some privileged epistemic status. But if his answer is to express his present opinion – which is surely what is requested – then this "momentary" shift in status must be guided by what his present opinion is. How does this guidance work?

One suggestion is that the respondent moves to a state of opinion derived from his own in two steps: (1) discarding beliefs so that the supposition receives more than minimal likelihood; (2) then (without further change in beliefs) regrading the alternatives left open so as to give the supposition maximal likelihood. This makes sense only if both steps are unambiguous. We can imagine a simple case. Suppose Peter has as "primary" beliefs A and B, and believes exactly what they jointly entail; he is asked to entertain the supposition $C - A$. In response he imaginatively moves into the epistemic position in which (1) B is the only primary belief, and (2) he assigns 0 to all alternatives left open by B which conflict with $(C - A)$ and then regrades the others in the same proportions as they had but with the maximum assigned to $(B \cap C - A)$.

This simple case already hinges on a certain hierarchical structure in Peter's opinion. Moreover it presupposes that those alternatives which were left open by B, but which conflict with his initial equally primary belief that A, had been graded proportionately as well. Even more structure must be present to guide the two steps in less simple cases. What if the beliefs had been, say, A, B, and D, and their joint consequences, and the supposition was compatible with each but not with the conjunction of any two? The discarding process can then be guided only if some hierarchy among the beliefs determines the selection.

Let us consider conditional personal probability as a possible means for describing structure of this sort. The intuitive Example 5.1 above about the mass of the moon is the sort often given to argue for the irreducibility of conditional probability. I could continue the example with: the mass of the moon seems to me to equally likely to be x as $(x + b)/2$, on the supposition that it is one of those two numbers. The two possibilities at issue here are represented by the degenerate intervals $[x]$, $[(x + b)/2]$, so both they and the supposition that one or other is the case (represented

by set $\{x, (x+b)/2\}$ their union) receive probability 0. The usual calculation of conditional probability, which would set $P(B \mid A)$ equal to $P(B \cap C|A)$ divided by $P(C|A \cap C)$, can therefore not be carried out. The suggestion that conditional probability is irreducible means that *two-place probability* $P(\ |\ )$ – probability of one thing *given* (*on supposition of*) another – is autonomous and cannot be defined in terms of the usual one-place ("absolute") probability. Rather the reverse: we should define $P(\ ) = P(\ |\ U)$, probability conditional on the tautology.

There is a good deal of literature on two-place ("irreducible conditional") probability (see Appendix). Despite many individual differences, general agreement concerning two-place probability extends to:

I. If P is a 2-place probability function then $P(-\mid A)$ is "normally" a (1-place) probability function with $P(A \mid A) = 1$.
II. These derivative 1-place probability functions [described in I.] are related at least by the *Multiplication Axiom:*

$$P(B \cap C|A) = P(B|A)\, P(C|A \cap C)$$

where A, B, C, ... are assumed to be in the domain and co-domain of the function. The "normally" restriction (eliminating at least $A = \Lambda$) is to be discussed below.

New non-trivial relationships between propositions are now definable. De Finetti suggested relations of local comparison of the following type:

$$A \text{ is } \textit{superior} \text{ to } B \text{ iff } P(A|A + B) = 1$$

where '+' marks exclusive disjunction: $A + B = (A - B) \cup (B - A)$.[3]

Example 5.2 Given any probability measure P it is easy to produce a 2-place function that has that character:

$$\text{define } P(A|B) = P(A \cap B)\,/P(B) \text{ if } P(B) > 0$$
$$= 1 \text{ if } P(B) = 0$$

That is a trivial 2-place function since it is definable from a 1-place function.

[3] De Finetti (1936). I want to thank John M. Vickers for bringing this to my attention; De Finetti's idea is developed considerably further, with special reference to zero relative frequency, in Vickers (1988), Sections 3.6 and 5.4. The relation here defined is slightly different from the so-named one in my (1979) – to which the name was somewhat better suited – for convenience in some of the proofs.

Example 5.3 Let U be the set of natural number $\{0, 1, 2, \ldots\}$. For index $n = 0, 1, 2, \ldots$ let p_n be the probability measure defined on all subsets of U by the condition that it assigns 0.1 to $\{x\}$ if x is in the set $\{10n, \ldots, 10n + 9\}$, and 0 otherwise. Define:

$$P(A|B) = p_n(A \cap B)/p_n(B) \text{ for the first index } n \text{ such that } p_n(B) > 0, \text{ if there is any such index; } = 1 \text{ otherwise.}$$

To verify the Multiplication Axiom note for instance that if $A \cap C$ is not empty, and $P(A \mid C) > 0$ then the first index n for which $p_n(A \cap C) > 0$ is the same as the first index m such that $p_m(C) > 0$. The "otherwise" clause will apply here only if $B = \Lambda$.

These examples are instances of the "lexicographic" probability models which I will discuss at some length below. We make the ideas of one- and two-place probability precise as follows.

A *space* is a couple $S = \langle U, F\rangle$ with U a non-empty set (the *worlds*) and F (the family of *propositions*) a sigma-field on U, that is:

(a) $U \in F$
(b) if $A, B \in F$ then $A - B \in F$
(c) if $\{A_i: i = 1, 2, \ldots\} \subseteq F$ then $\cup \{A_i\} \in F$

A (1-place) *probability measure* P on space $S = \langle U, F\rangle$ is a function mapping F into the real numbers, subject to

1. $0 = P(\Lambda) \leq P(A) \leq P(U) = 1$
2. $P(A \cup B) + P(A \cap B) = P(A) + P(B)$ (*finite additivity*)
3. If $E_1 \subseteq E_2 \subseteq \ldots \subseteq E_n \subseteq \ldots$ has union E, then $P(E) = \sup\{P(E_n): n = 1, 2, \ldots\}$ (*continuity*)

Property 3 is in this context equivalent to *countable additivity*:

4. If $\{E_n: n = 1, 2, \ldots\}$ are disjoint, with union E, then $P(E) = \Sigma\{P(E_n): n = 1, 2, \ldots\}$

and also to the dual continuity condition for countable intersection. The general class of two-place probability measures to be defined now will below be seen to contain a rich variety of non-trivial examples.

A *2-place probability measure* $P(- \mid -)$ on space $S = \langle U, F\rangle$ is a map of $F \times F$ into real numbers such that

I. (Reduction Axiom) The function $P(- \mid A)$ is either a probability measure on S or else has constant value $= 1$.
II. (Multiplication Axiom)

$$P(B \cap C|A) = P(B|A)\, P(C|B \cap A)$$

for all A, B, C, in F.

If $P(|\ A)$ is a (1-place) probability measure, I shall call A *normal* (for P), and otherwise *abnormal*. ("Absurd" might have been a better name; it is clearly a notion allied to self-contradiction.) The definition of 2-place probability allows for the totally abnormal state of opinion ($P(A \mid B) = 1$ for all A and B). It should not be excluded formally, but I shall tacitly exclude it during informal discussion. Here are some initial consequences of the definition. The variables range of course over propositions (members of family F in space $S = \langle U, F \rangle$).

[T2.1] $P(X|A) = P(X \cap A|A)$
[T2.2] If A is normal, so are its supersets
[T2.3] If A is abnormal, so are its subsets
[T2.4] B is abnormal iff $P(-B \mid B) = 1$; iff $P(B \mid A) = 0$ for all normal A.

Let us call the case in which only Λ is abnormal the "Very Fine" case; that there are Very Fine 2-place probability measures on infinite fields will follow from results below.

We add one consequence which is related to De Finetti's notion of conglomerability:

[T2.5] Condition Continuity: If $\{E_n\}$ is a countable increasing chain – i.e. E_n part of E_{n+1} – with union E, and $P(E_n \mid E) > 0$ for all n, then $P(X \mid E)$ is the limit of the numbers $\{P(X|E_n)\}$.

To prove this: $P(X|E_n) = P(X \cap E_n|E)\ /P(E_n \cap E)$, so since E is normal, this follows from the principle of condition continuity which can be demonstrated for (one-place) probability measures.[4]

The Idea of the *A Priori*

In any conception of our epistemic state there will be propositions which are not epistemically distinguishable from the tautology U – let us say these are *a priori* for the person. This notion is the opposite of the idea of abnormality:

$$\text{A is } a \ \textit{priori for } P \ \text{ iff } P(A|X) = 1 \ \text{ for all } X, \text{iff } U - A \ \text{ is abnormal for } P.$$

What is *a priori* for a person is therefore exactly what is certain for him or her on any supposition whatsoever. This notion generalizes unconditional certainty, i.e. $P(A) = 1$. The strongest unconditional probability equivalence relation between A and B is that their symmetric difference $(A + B)$ has measure zero. We can

[4] See B. De Finetti (1972), Section 5.22.

generalize this similarly. As our strictest epistemic equivalence relation between two propositions we have *a priori* equivalence (their symmetric difference has probability 0 on all normal suppositions):

$$A \langle P \rangle B \text{ iff } A + B \text{ is abnormal.}^5$$

The abnormal propositions are the ones *a priori* equivalent to the empty set (the self-contradiction) and the *a prioris* are the ones *a priori* equivalent to the tautology. (Of course these are subjective notions: we are speaking of what is *a priori* for the person with this state of opinion.)

Note now that $A\langle P\rangle B$ iff $P(A \mid A + B) = 1$ and $P(B \mid A + B) = 1$, since additivity would not allow that if $A + B$ were normal. We can divide this equivalence relation into its two conjuncts:

Definition A $P > B$ iff $P(A \mid A + B) = 1$.

This is the relationship of "superiority" mentioned above.

[T3.1] If A logically implies B then $B\ P > A$.

It follows from the definition that $A\ P > A$. In a later section I shall also show that $P >$ is transitive. Clearly if $A + B$ is normal, then $A\ P > B$ means that A is comparatively superior to B, in the sense that A is certainly true and B certainly false, on the supposition that one but not both are the case. But if $A + B$ is abnormal then the relationship $A\ P > B$ amounts to $A\langle P\rangle B$. The right reading for "$P >$" is therefore "is superior to or *a priori* equivalent to". To be brief, however, I'll just say "A is superior to B" for "$A\ P > B$", and ask you to keep the qualifications in mind.

Full Belief Revisited

The beliefs I hold so strongly that they are *a priori* for me are those whose contraries are all abnormal. There is a weaker condition a proposition K can satisfy: namely that any normal proposition which implies K is superior to any that are contrary to K. Consider the following conditions and definitions:

(A1)	Normality:	K is normal
(A2)	Superiority:	If A is a non-empty subset of K while B and K are disjoint, then $A\ P > B$
(A3)	Contingency:	The complement $U - K$ of K is normal.

We can restate the "Superiority" condition informally as follows:

[5] From this point on I shall drop the ubiquitous "for P" unless confusion threatens, and just write "a priori", "abnormal", etc. leaving the context to specify the relevant 2-place probability measure.

Superiority: the alternatives K leaves open are all superior to any alternative that K excludes.

We can deduce from these conditions something reminiscent of Carnap's "regularity" (or Shimony's "strict coherence"):

(A4) Finesse: all non-empty subsets of K are normal.

Definition K is a *belief core* (for P) iff K satisfies (A1)–(A3).

Note that the *a priori* propositions satisfy (A1) and (A2), though definitely not (A3), but rather its opposite. However, the following elementary results show that all the *a prioris* are among the propositions implied by belief cores.

[T4.1] If K is a belief core then $P(K \mid U) = 1$.
[T4.2] If K is a belief core then $A\ P > K$ iff K implies A.
[T4.3] If K is a belief core and A is *a priori* then K is a subset of A.

To characterize the full beliefs we need take into account the extreme possibility of there being no belief cores; in that case we still want the *a prioris* to be full beliefs of course. (This corresponds to what I have elsewhere called "Zen minds": states of opinion in which nothing is fully believed if it is subjectively possible to withhold belief.) In view of the above we have several equivalent candidates for this characterization, of which we can choose one as definition:

Definition A is a *full belief* (for P) iff (i) there is a belief core, and $A\ P > K$ for some belief core K; or (ii) there is no belief core, and A is *a priori*.

[T4.4] The following conditions are equivalent:
(a) A is a full belief (for P).
(b) Some proposition J which is either *a priori* or a full belief core is such that $A\ P > J$.
(c) A is implied either by an *a priori* or by a belief core (for P).

Very little in this discussion of full belief hinges on the peculiarities of probability. Indeed, (conditional) probability enters here only to give us intelligible, non-trivial resources for defining the notions of subjective superiority and a-prioricity (equivalently, [ab]normality). If those notions could be acceptably primitive, the account of full belief here would just take the following form:

A belief core is a proposition K such that:

(a) K and its complement are both normal; (b) K does not leave open any abnormal alternatives; (c) any alternatives left open by K are superior to any alternatives K excludes.

A belief is any proposition implied by a belief core (or *a priori*).

As before, a proposition is here called an alternative left open by K exactly if it is non-empty and implies K.

Identification of Full Beliefs

By the definition I gave of full belief, beliefs are clustered: each belongs to a family $\{A: A\ P > K\}$ for some belief core K (if there are any at all), which by [T4.2] sums up that cluster exactly:

$$K \text{ is the intersection of } \{A: A\ P > K\}$$
$$\{A: A\ P > K\} = \{A: K \text{ implies } A\}.$$

We can now prove that these clusters form a chain, linearly ordered by set inclusion (implication).

[T5.1] If K, K' are belief cores, then either K is a subset of K' or K' is a subset of K.

For the proof let $A = K - K'$ and $B = K' - K$. Each is normal if non-empty, by (A4). If B is not empty then, since A is disjoint from K', it follows by (A2) that $B\ P > A$. By parity of reasoning, if A is not empty, then $A\ P > B$. But we cannot have both unless $A + B$ is abnormal, and hence empty by [T2.2] and (A4). So we conclude that either A or B or both are empty; hence at least one of K and K' is a subset of the other.

This result is crucial for the characterization of full belief. It is therefore worthwhile to note that the only ingredients needed for the proof were the features of Superiority and Finesse of belief cores, plus the following characteristic of the superiority relationship: if $A\ P > B$ and $B\ P > A$ then $A + B$ is abnormal. Here is an illustrative example:

Example 5.4 For the center of mass of the remains of Noah's ark, Petra has subjective probability 1 for each of the following three propositions: that it lies in the Northern Hemisphere (which part of the earth includes the Equator), that it lies in the Western Hemisphere, and thirdly that it lies North of the Equator. But only the first two of these propositions are among her full beliefs; the third is not. On the supposition that one but only one of these beliefs is true, she gives 100 % probability to the first proposition, that it lies in the Northern Hemisphere.

Note that the last sentence implies that the first proposition is superior to the second, although both are full beliefs. I will give a formal reconstruction of this example later on.

Writing K^* for the intersection of all the belief cores, we conclude that if A is a full belief, then K^* implies A. But is K^* itself a belief core? Does it have 100 % probability? Is it even non-empty? This is the problem of transfinite consistency of full belief in our new setting.

[T5.2] The intersection of a non-empty countable family of belief cores is a belief core.

For proof, assume there is at least one belief core; call it K. Assume also that the belief cores are countable and form a chain (the latter by [T5.1]), and call the intersection K^*. Countable additivity of the ordinary probability measure $P() = P(|U)$ is equivalent to just the right continuity condition needed here: the probabilities of the members of a countable chain of sets converge to the probability of its intersection. Since in our case all those numbers equal 1, so does $P(K^*)$. Therefore also K^* is not empty, and thus normal because it is a subset of at least one belief core. Moreover, so are its non-empty subsets, so that they are normal too. Its complement $U - K^*$ includes $U - K$, and is therefore also normal.

We have now seen that K^* satisfies conditions (A1), (A3), and (A4), and need still to establish (A2). If A is a normal subset of K^*, and hence of all belief cores, and B is disjoint of K^*, we have $P(A \mid A + (B - K')) = 1$ for all belief cores K'. But the sets $B - K'$ form an increasing chain whose union is $B - K^* = B$. Hence also the sets $A + (B - K')$ here form such a chain with union $A + B$. To conclude now that $P(A \mid A + B) = 1$, we appeal to [T2.5], the principle of Condition Continuity. This ends the proof.

The significance of this result may be challenged by noting that the intersection of countably many sets of measure 1 also has measure 1. So how have we made progress with the transfinite lottery paradox? In four ways. The first is that in the representation of opinion we may have a "small" family of belief cores even if probability is continuous and there are uncountably many propositions with probability 1. The second is that no matter how large a chain is, its intersection is one of its members if it has a first (= "smallest") element. The third is that the following is a condition typically met in spaces on which probabilities are defined even in the most scientifically sophisticated applications:

(*) Any chain of propositions, linearly ordered by set inclusion, has a countable subchain with the same intersection.

[T5.3] If (*) holds and there is at least one belief core, then the intersection of all belief cores is also a belief core.

This is a corollary to [T5.2].

Fourthly, farther below I will also describe an especially nice class of models of fine-grained opinion for which we can prove that the intersection of the belief cores, if any, is always also a belief core ("lexicographic probability"). There are no countability restrictions there.

To What Extent Does Belief Guide Opinion?

It is not to be expected that every two-place probability function is admissible as a representation of (possible) opinion. If we want to use this theory in descriptive epistemology, it is necessary to look for kinds of probability functions that have interesting structure. There are models in which there are no belief cores at all. Combining our previous Examples 5.1 and 5.2, take Lebesgue measure m on the

unit interval, and trivially extend it to a two-place function by $P(A \mid B) = m(A \cap B)/m(B)$ if defined and $P(A \mid B) = 1$ if not (though A, B in domain of m). Then every unit set $\{x\}$ is in the domain and is abnormal. Therefore there is no set all of whose subsets are normal, and hence no belief cores. (The absence of belief cores in our present example derives from its triviality, and not from the continuity.) Obviously then, if this represents someone's opinion, his opinions are not guided or constrained by his beliefs (which include only the *a priori*).[6]

At the other extreme from this example, there is the Very Fine case of a probability function P for which every non-empty set is normal.

Definition P is *belief covered* if the union of the belief cores equals U.

In that case, P is Very Fine. For let A be any non-empty proposition; there will be some belief core K such that $K \cap A$ is not empty, hence normal, thus making A normal.

Example 5.3 furnishes us with a relatively simple example of this sort. Recall that P is there constructed from the series $p_0, p_1, \ldots, p_n, \ldots$ where the whole probability mass of p_n is concentrated (and evenly distributed) on the natural numbers $\{10n, \ldots, 10n + 9\}$. In this example, the belief cores are exactly the sets

$$K_0 = \{0, \ldots, 9\}, K_1 = \{0, \ldots, 19\},$$
$$K_2 = \{0, \ldots, 29\}, \ldots K_i = \{0, \ldots, 10i + 9\}$$

Clearly K_i is normal, since $P(-|K_i) = P(-|K_0) = P_0$. The complement of K_i is normal too, for

$$P(-|U - K_i) = P(-|\{10(i + 1), \ldots\}) = p_{i+1}.$$

If A is a non-empty subset of K_i and B is disjoint from K_i, then A is superior to B. Specifically, the first n such that $p_n(A) > 0$ can be no higher than i in that case, while the first m such that $p_m(B) > 0$ can be no lower than $i + 1$. Therefore, the first k such that $p_k(A + B) > 0$ will assign a positive probability to A and *zero* to B.

These belief cores clearly cover U; P is belief covered and Very Fine. Indeed, the belief cores are well-ordered,

Define the *belief remnants*

$$R_0 = K_0$$
$$R_{j+1} = K_{j+1} - K_j \, (j = 0, 1, 2, \ldots).$$

[6]Could a person's opinion be devoid of belief cores? Our definitions allow this, and it seems to me this case is related to the idea of a "Zen mind" which I have explored elsewhere (van Fraassen 1988).

Clearly $p_i = P(|\ R_i)$; for example, $p_1 = P(|\ \{10, \ldots, 19\})$. Probabilities conditional on belief remnants (beliefs remaining upon retrenchment to a weaker core) determine all probabilities in this case:

$$P\,(-|A) = P\,(-|A \cap R_i) \text{ for the first } i \text{ such that } P\,(A|R_i) > 0.$$

This says quite clearly that (in this case) belief guides opinion, for probabilities conditional on belief remnants are, so to speak, all the conditional probabilities there are.

The Multiplication Axiom Visualized

In the basic theory of two-place probability, the Multiplication Axiom places the only constraint on how the one-place functions $P(|\ A)$, $P(|\ B)$, ... are related to each other. It entails that the proposition $A - B$ is irrelevant to the value of $P(B \mid A)$ – that this value is the same as $P(A \cap B \mid A)$ – and that the usual ratio-formula calculates the conditional probability when applicable. Indeed, the ratio formula applies in the generalized form summarized in the following:

If X is a subset of A which is a subset of B, then:

[T7.1] if $P(A \mid B) > 0$ then $P\,(X|A) = P\,(X|B) : P\,(A|B)$

[T7.2] if X is normal, then $P\,(X|B) \leq P\,(X|A)$.

There is another way to sum up how the Multiplication Axiom constrains the relation between $P(|\ A)$ and $P(|\ B)$ in general. When we consider the two conditional probabilities thus assigned to any proposition that implies both A and B, we find a proportionality factor, which is constant when defined.

[T7.3] If $P(A \mid B) > 0$ then there is a constant $k \geq 0$ such that for all subsets X of $A \cap B$, $P\,(X|A) = kP\,(X|B)$. The constant $k = k\,(A, B) = [P\,(B|A)\ /\ P\,(A|B)]$, defined provided $P(A \mid \mathrm{B}) > 0$.

Equivalence Relations on Propositions

In the main literature on two-place probability we find an equivalence relationship other than *a priori equivalence*, which I shall call *surface equivalence*:[7]

[7] In earlier treatments of two-place probability this relationship has appeared as a special axiom: If $P(A \mid B) = P(B \mid A) = 1$ then $P(\cdot \mid A) = P(\cdot \mid B)$.

[T8.1] The following conditions are equivalent:
(a) $P(\cdot|A) = P(\cdot|B)$
(b) $P(A + B|A \cup B) = 0$ or else both A, B are abnormal
(c) $P(A|B) = P(B|A) = 1$.

(I use the dot for function notation: (a) means that $P(X|A) = P(X|B)$ for all X.) It is easy to prove that (a) implies (b), for if (a) and either A or B is normal then both are normal. Secondly suppose (b). If A, B abnormal then (c) follows. Suppose then that one of A, B is normal, so $A \cup B$ is normal. Then either $P(A|A \cup B)$ or $P(B|A \cup B)$ is positive; suppose the first. Since $A \cap B \subseteq A \subseteq A \cup B$ it follows by [T7.1] that $P(B \mid A) = P(B \cap A \mid A \cup B) : P(A \mid A \cup B)$. But $P(A - B|A \cup B) = 0$ so $P(A \cap B|A \cup B) = P(A|A \cup B) > 0$; hence we conclude that $P(B \mid A) = 1$. Accordingly, B too is normal and $P(B|A \cup B)$ is positive; the same argument leads *mutatis mutandis* to $P(A \mid B) = 1$. Therefore (b) implies (c).

Finally suppose (c). If A is abnormal, then so is B and (a) follows at once. If A is normal, then B and $A \cup B$ are also normal. But then $P(A \cap B|A \cup B) = P(A|A \cup B) = P(B|A \cup B) = 1$, using (c) and the Multiplication Axioms. Hence $P(X \cap A \cap B|A \cup B) = P(X \cap A|A \cup B) = P(X|A)$ and by similar reasoning $P(X \cap A \cap B|A \cup B) = P(X|B)$. Therefore (c) implies (a).

The relationship of *a priori* equivalence is much "deeper". As prelude let us introduce another operation on probability functions which is something like "deep conditionalization". Instead of raising a proposition to the status of subjective certainty it raises it to subjective aprioricity. To prevent confusion, I shall call this "relativization".[8]

Definition The *relativization of P to A* is the function $P//A$ defined by $P//A(X \mid Y) = P(X \mid Y \cap A)$ for all X, Y.[9]

[T8.2] The following are equivalent:

(i) $P(A\|\cdot) = P(B\|\cdot)$	
(ii) $P(-\|\cdot \cap A) = P(-\|\cdot \cap B)$	(i.e. $P//A = P//B$)
(iii) $P(A\|A + B) = P(B\|A + B) = 1$	(*a priori* equivalence)
(iv) $A + B$ is abnormal	(i. e. $A\langle P\rangle B$)

(For proof that (ii) implies (iv), use the *Lemma*: if A, B are disjoint and abnormal then $A \cup B$ is abnormal.).

[T8.3] *A priori* equivalence implies surface equivalence.

[8] As I have argued elsewhere (van Fraassen 1981a) this construction provides us with the "right" clue to the treatment of quantification and of intuitionistic implication in so-called probabilistic (or generally, subjective) semantics.

[9] In my (1981a), $P//A$ was designated as P^A and called "P conditioned on A." I now think this terminology likely to result in confusion, and prefer "P relativized to A."

The converse does not hold, as can be seen from our Example 5.3 in Section 2.1, where U is the set of natural numbers, and is surface equivalent, but not *a priori* equivalent, to $\{0, 1, \ldots, 9\}$. For $P(\{0, 1, \ldots, 9\} \mid \{10, \ldots, 19\}) = 0$ there.

Implication Relations; Superiority Is Transitive

We are representing propositions by means of a field of sets, whose elements are thought of as alternative possible situations or worlds. Accordingly, "A implies B" can be equated only with "$A \subseteq B$." But when two propositions are *a priori* equivalent for P then they are not distinguishable as far as P is concerned. Therefore we can introduce a less sensitive partial ordering as a "coarser" implication relationship with respect to a given two-place probability measure.

[T9.1] The following are equivalent:
(a) A *P-implies* B: $P(A \mid \cdot) \leq P(B \mid \cdot)$
(b) $A - B$ is abnormal
(c) for all X, if $P(A \mid X) = 1$ then $P(B \mid X) = 1$.

The superiority relation is not a (converse) implication relationship, despite formal similarities. If A is superior to B, A may still have probability *zero* conditional on B, for example. It is just that the supposition that A has to be given up – in fact, denied – before B comes into play in our thinking. The hierarchy so indicated John Vickers (*loc. cit.*) calls the De Finetti hierarchy. As he points out, it is crucial to this role that we can describe the comparison in terms of a transitive relationship.

In fact, only one further point needs to be made concerning normality to show that the propositions form a partially ordered set under superiority (with the abnormal propositions forming an equivalence class at the bottom of this p.o.set).

[T9.2] If X is a subset of Y and Y a subset of normal set E, and $P(Y \mid E) > 0$, then $P(X \mid Y) = 0$ iff $P(X \mid E) = 0$.

[T9.3] $P >$ is transitive.

Let it be given that $A \ P > B$ and $B \ P > C$; we need to prove that $A \ P > C$. To visualize the proof, think of a Venn Diagram with the following labels for relevant propositions:

$$1 = (A - B - C), \quad 2 = (AB - C), 3 = B - A - C,$$

$$4 = AC - B$$

$$5 = BC - A, 6 = C - B - A$$

$$E = (A + B) \cup (B + C) = (A + B) \cup (B + C) \cup (A + C)$$

I will denote unions by concatenation of the labels; thus $E = 123456$ and $A - C = 12$. We now consider all possible cases.

(1) $A + C$ is abnormal. Then $P(A \mid A + C) = 1$
(2) $A + C$ is normal; then also E is normal.

Hence $P(A + B \mid E)$ or $P(B + C \mid E)$ or both are positive; we proceed to the possible subcases:

(2.1) $P(A + B \mid E) > 0$. By the given and [T9.2] it follows that $P(B - A \mid E) = 0$, i.e. $P(3 \mid E) = P(5 \mid E) = 0$.
(2.11) Assume $P(B + C \mid E) > 0$. By [T9.2], and the given, also $P(C - B \mid E) = 0$, so $P(4 \mid E) = P(6 \mid E) = 0$. Altogether, $P(3456 \mid E) = 0$ hence $P(12 \mid E) = P(A - C \mid E) = 1$. It follows that $P(A + C \mid E) > 0$, so by [T9.2] $P(56 \mid A + C) = P(C \mid A + C) = 0$, and therefore $P(A \mid A + C) = 1$.
(2.12) Assume $P(B + C \mid E) = 0$. Then $P(2346 \mid E) = 0$, so altogether $P(23456 \mid E) = 0$. Hence $P(1 \mid E) = 1$. It follows that $P(A + C \mid E) > 0$, therefore by [T9.2] again $P(56 \mid A + C) = 0$. It follows then that $P(A \mid A + C) = 1$.
(2.2) $P(A + B \mid E) = 0$ and $P(B + C \mid E) > 0$. The former entails that $P(1345 \mid E) = 0$. The latter entails by [T9.2] that $P(C - B \mid E) = P(46 \mid E) = 0$. Altogether therefore $P(13456 \mid E) = 0$ and $P(2 \mid E) = 1$. Therefore $P(A + C \mid E) > 0$, and so $P(C - A \mid A + C) = 0$ by [T9.2]; therefore $P(A \mid A + C) = 1$. This ends the proof.

Adding this to the fact that $P >$ is reflexive, we conclude that $P >$ is a partial ordering of the field of propositions. The abnormal propositions form a $\langle P \rangle$ equivalence class at the very bottom of this partial ordering.

A Large Class of Models

I will define a class of models such that P satisfies principles I–II iff P can be represented by one of these models, in the way to be explained. The class will be chosen large; a special subclass ("lexicographic models") will yield nontrivial, easily constructed examples to be used in illustrations and refutations. (The term "lexicographic" is used similarly in decision theory literature; see Blume et al. 1991a, b.)

A model begins with a sample space $S = \langle U, F \rangle$, where U is a non-empty set (the universe of possibilities) and F a sigma-field of sets on U (the propositions). We define the subfields:

$$\text{if } A \text{ is in } F \text{ then } FA = \{E \cap A : E \text{ in } F\};$$

thus FA is a field on A. For each such field designate as PA the set of probability measures defined on FA. (When A is empty, $FA = \{A\}$ and PA is empty.) The restriction of a member p of PA to a subfield FB, with B a subset of A, will be designed $p \mid FB$. Finally let PS be the union of all the sets PA, A in F.

A model M will consist of a sample space S as above, and a function π defined on a subset of F, with range in PS. That is, π associates some probability measure on some subfield with certain propositions. (These will be the normal propositions.) I will abbreviate "$\pi(A)$" to "πA", and when p is in FB I will also designate B as Up (the universe of p). Thus $B = U\pi A$ means that π associates with A a probability measure defined on the measurable subsets of B, i.e. on the propositions which imply B, i.e. on FB. The function π is subject to the following conditions:

(M1) $\pi A(A)$ is defined and positive.
(M2) If $\pi B(A)$ is defined and positive, then πA is defined
(M3) If $\pi B(A)$ is defined and positive, then $\pi A \mid F(A \cap B)$ is proportional to $\pi B \mid F(A \cap B)$.

This does not entail that if $\pi B(A \cap B) > 0$ then $\pi A(A \cap B) > 0$, because the proportionality constant can be 0 (in which case πA gives 0 to all members of $F(A \cap B)$ – see further the discussion of [T7.3] which suggested this condition). It is easy to see what the constant of proportionality has to be:

[T10.1] If $\pi B(A)$ is defined and positive, then
$\pi A \,|F\,(A \cap B) : \pi B|\,F\,(A \cap B)$
$= \pi A\,(A \cap B) : \pi B\,(A \cap B)$.

Finally we define what it means for one of these functions to represent a two-place function:

Definition Model $M = \langle S,\ \pi\rangle$ with $S = \langle U,\ F\rangle$ represents binary function P iff the domain of P is F and for all A, B in F, $P\,(A|B) = \pi B\,(A \cap B)\,/\,\pi B(B)$ if defined, and $= 1$ otherwise.

It is easy to prove that:

[T10.2] If P is represented by a model, then P is a two-place probability measure.

Conversely, suppose that P is a two-place probability measure in the sense of satisfying I–II, defined on F in space $S = \langle U,\ F\rangle$. For all normal sets A of P define πA on FA by:

$$\pi A(B) = P\,(B|A)\,.$$

That (M1) and (M2) are satisfied by $M = \langle S, \pi\rangle$ follows at once from the facts about normal sets. Suppose now, equivalently to the antecedent of (M3) that A and B are normal sets with $\pi(B) > 0$. To prove that (M3) holds, suppose that $\pi B(A)$ is defined and positive, so that B and A are normal sets, $P(A \mid B) > 0$. Then according to [T10.3], for each subset X of $A \cap B$ we have $P\,(X|A) = [P\,(B|A)\,/P\,(A|B)]\,P\,(X|B)$. Therefore here $\pi A(X) = [P(B \mid A)/P(A \mid B)]\pi B(X)$. In conclusion:

[T10.3] If P is a two-place probability measure satisfying the principles I–II, then P is represented by a model.

Having established this representation result, we now look for easily constructed models, for illustration, refutation of conjectures, and exploration of examples.

Definition Model $M = \langle S, \pi \rangle$ with $S = \langle U, F \rangle$ is *lexicographic* iff there is a sequence (well-ordered class) SEQ of 1-place probability measures defined on the whole of F, such that $\pi B(A) = q(A \cap B)/q(B)$ for the first member q of the sequence SEQ such that $q(B) > 0$; πB is undefined when there is no such q.

The members of SEQ correspond to the probabilities conditional on *belief remnants* (see discussion in Section 6 of Example 5.3). We will say that πA comes before πB in SEQ exactly when the first q in SEQ such that $q(A) > 0$ comes before the first q in SEQ such that $q(B) > 0$. It is easily checked that $M = \langle S, \pi \rangle$ is a model. Specifically, if A is a subset of B then πB will not come after πA, since whatever measure assigns a positive value to A will then assign one to B. Neither can πA come after πB if $\pi B(A) > 0$; in that case $\pi A = \pi B$. Consequently condition (M3) is easily verified: the proportionality constant = 1.

It is now very easy to make up examples of 2-place probability measures. Just take two or three or indeed any number, finite or infinite, of ordinary probability measures and well-order them. A special example, whose existence depends on the axiom of choice is this: let SEQ contain all one-place probability measures defined on given domain F. In that case, the only abnormal proposition is the empty set (the self-contradiction). Also the only *a priori* is the tautology. Short of this, we could of course have a sequence which does not contain literally all the definable probability measures, but contains all those which give 1 to a given set A. In that case, all propositions other than Λ that imply A are normal. Let us call P *Very Fine on* A in such a case. (The case of P Very Fine on U was already called "Very Fine" above.) Note that one of the defining conditions of a belief core K was that P had to be Very Fine on K.

Belief in a Lexicographic Model

I will first show that in a lexicographic model, the intersection of all belief cores, if any, is always a belief core too. Since this does not depend on cardinality or the character of the sample space, the result adds significantly to the previous theorems. Then I will construct a lexicographic model to show that in general not all propositions with probability 1 are full beliefs. This model will be a reconstruction of Example 5.4 (Petra and Noah's Arc).

[T11.1] If P is represented by lexicographic model $M = \langle S, p\rangle$ defined by w.o. sequence SEQ, and A, B are disjoint normal sets for P, then the following are equivalent:
(i) $A\ P{>}B$ and not $B\ P{>}A$
(ii) πA comes before πB in SEQ.

[T11.2] If P is represented by lexicographic model $M = \langle S, \pi\rangle$ defined by w.o. sequence SEQ, and K, $K1$, $K2$ are belief cores with K a proper subset of $K1$ and $K1$ a proper subset of $K2$ then $\pi\ (K1 - K)$ comes before $\pi(K2 - K1)$ in SEQ

For proof note that $K1 - K$ and $K2 - K1$ are disjoint normal sets (*belief remnants*), so [T11.1] applies.

[T11.3] If P is represented by lexicographic model $M = \langle S, \pi\rangle$ defined by w.o. sequence SEQ, then the intersection of all its belief cores is also a belief core.

For convenience in the proof, call $\pi(K'' - K')$ a marker of K' when both are belief cores and K' is a proper subset of K''. This measure exists in the model since the set is normal. If no superset of K is a core, call $\pi(U - K)$ its marker. We now consider the whole class of markers; it must have a first member in SEQ. Let that be $p^* = \pi(K2 - K1)$. It remains to prove that K1 is the intersection of all the belief cores. Suppose to the contrary that core K is a proper subset of $K1$. Then by [T11.2] marker $p\ (K1 - K)$ comes before p^* in SEQ – contra the hypothesis. This ends the proof.

At this point we know enough about lexicographic models in general to exploit their illustrative uses. Recall Example 5.4. Petra has subjective probability 1 for N: the center of mass of Noah's Ark lies in the Northern Hemisphere and also for W: that it lies in the Western Hemisphere. I shall take it here that the Equator is part of the Northern Hemisphere; she has probability 1 that –EQ: it does not lie on the Equator. Let me add here that she also has probability 1 that –GR: it does not lie on the Greenwich Meridian (which I shall here take to be part of the Western Hemisphere). But she has probability 1 that it lies in Northern Hemisphere on the supposition that $N + W$: it lies either in the Northern or in he Western Hemisphere but not both (which supposition has 0 probability for her).

For the sample space, let U be the entire surface are of the Earth, and let F be the family of measurable subsets of U in the usual sense, so we can speak of area and length where appropriate. Let us first define some measures and classes of measures:

$$mA(X) = \text{area of } X \cap A : \text{area of } A$$
$$1A(X) = \text{length of } X \cap A : \text{length of } A$$

where XX is a subset of F

$$M(XX) = \text{the class of all measures on } F$$

which give 1 to a member of XX.

The sequence SEQ is now pieced together from some other well ordered sequences as follows:

$$\text{SEQ} = m\,(N \cap W)\,, l\,(\text{GR} \cap N)\,, l\,(\text{EQ} \cap W)\,, \text{S1}, \text{S2}, mN, mW, mU, \text{S3}, \text{S4}$$

where the indicated subsequences are:

S1: a well-ordering of $\{1A$: A is a subset of $N \cap W$ with 0 area but positive length$\}$,
S2: a well-ordering of $M\{A$ in F: A is a non-empty subset of $N \cap W$ with 0 area and 0 length$\}$,
S3: a well-ordering of $\{1A$: A is has 0 area but positive length$\}$,
S4: a well-ordering of $M\{A$ in F: A is a non-empty and has 0 area and 0 length$\}$.

Let us call the so constructed lexicographic model PETRA, in her honor. The construction is a little redundant: we would be constructing a Very Fine measure already if we just took the tail end mU, S3, S4. Preceding them with the others has the effect of establishing desired superiority relationships. For example, $N + W$ first receives a positive value from mN, which gives 0 to $W - N$, so that $P(N \mid N + W) = 1$. I made GReenwich similarly superior to EQuator.

[T11.4] $N \cap W$ is a belief core in PETRA.
[T11.5] No proper subset of $N \cap W - \text{GR}$ is a belief core.

For let X be such a subset; to be a belief core it must have probability 1 tout court, so its area equals that of $N \cap W$. Let X' be one of its non-empty subsets with 0 area. Then X′ as well as X itself are disjoint of $N \cap W \cap \text{GR} = Y$. The first measure to assign positive value to $X' + Y$ is the second member of SEQ, namely $m(\text{GR} \cap N)$, which assigns 1 to Y (because GR is part of W) and 0 to X'. Therefore X' is not superior to Y, and so X is not a belief core.

[T11.6] In PETRA some propositions not among its full beliefs have probability 1.

In view of the preceding, it suffices to reflect that $N \cap W - \text{GR}$ has proper subsets with probability 1; for example $N \cap W - \text{GR} - \text{EQ}$.

Appendix

A1. Previous Literature

The basic theory of two-place probability functions is a common part of a number of theories. Such probability functions have been called *Popper functions* because Popper's axioms originally presented in his *The Logic of Scientific Discovery* (1959) were adopted by other writers (see Harper 1976; Field 1977; van Fraassen 1979, 1981a, b). Carnap used essentially the same axioms for his "c-functions", but

concentrated his research on those which derive trivially from one-place probability functions ("m-functions"). Reichenbach's probability was also irreducibly two-place. I have mentioned De Finetti's paper (1936) which introduced the idea of local comparisons (like my "superior"; Vickers' "thinner"); see also Section 4.18 in his *Theory of Probability*, vol. 1. The most extensive work on two-place probability theory is by Renyi (1955, 1970a,b). The theory of two-place probability here presented is essentially as explored in my (1979), but with considerable improvement in the characterization of the described classes of models. Finally, the discussion of supposition in section "Supposition and two-place probability" is related to work on belief revision, much of it indebted to ideas of Isaac Levi; see Gardenfors 1988 for a qualitative version.

A2. Transfinite Consistency

The ordering $P(A) \leq P(B)$ extends the partial ordering of logical implication: if $A \subseteq B$ then $P(A) \leq P(B)$. Unfortunately, the ordering $P(A) < P(B)$ does not extend in general the partial ordering of proper implication: $P(A) = P(B)$ is possible even when $A \neq B$. Indeed, this is inevitable if there are more than countably many disjoint propositions. As a corollary, the intersection of all propositions of maximal probability may itself even be empty. Kolmogoroff himself reacted to this problem by suggesting that we focus on probability algebras: algebras of propositions reduced by the relation of equivalence *modulo* differences of measure *zero*: $P(A + B) = 0$. (See Birkhoff (1967), XI, 5 and Kappos (1969), II, 4 and III, 3.)

The difficulty with this approach is that a probability algebra does not have the structure usually demanded of an algebra of propositions. For the latter, the notion of truth is relevant, so it should be possible to map the algebra homomorphically into {0, 1}. As example take the unit interval with Lebesgue measure, reduced by the above equivalence relation. This is a probability (sigma-)algebra. Let T be the class of elements designated as true, i.e. mapped into 1, and let A with measure x be in T. Then A is the join of two disjoint elements of measure $x/2$ each. Since the mapping is a homomorphism, one of these is in T. We conclude that T contains a countable downward chain $A_1, A_2, \ldots$ with the measures converging to *zero*. Therefore its meet is the *zero* element of the algebra. The meet should be in T because it is the countable meet of a family of "true" propositions; but it can't be in T, since the *zero* element is mapped into 0.

This "transfinite inconsistency" of the propositions which have probability *one*, was forcefully advanced by Patrick Maher (1990) as a difficulty for the integration of subjective probability and belief. My conclusion, contrary to Maher's, is that the role of subjective probability is to grade the alternatives left open by full belief. That automatically bestows maximal probability on the full beliefs, but allows for other propositions to also be maximally probable. The question became then: how are the two classes of maximally probable propositions to be distinguished?

A3. Rejection of the Bayesian Paradigm?

While I hold to the *probabilist* conviction that our opinion is to be represented by means of probability models, I reject many features associated with so-called "Bayesian" views in epistemology. In the present context, a main difference concerns the status of probability *one*. Conditionalization of an absolute (one-place) probability function cannot lower probability from *one* nor raise it from *zero*. As a result, such changes have often been relegated to epistemological catastrophes or irrational shifts of opinion. This is definitely not so in all probabilist work in epistemology (Isaac Levi and William Harper provide notable exceptions). In my view, probability *one* is easily bestowed, and as easily retracted, especially when it is *only* maximal unconditional probability (conditional on the tautology).

Obviously, then, I reject the naive Pascalian equation that a bet on A, with any payoff whatsoever, is worth to me my probability for A times that payoff. I think that Pascal's equation holds under restricted circumstances, with relevant assumptions kept fixed and in place. I mean this roughly in the sense of the "constructivist" view of subjective probability suggested in various ways by Glenn Shafer and Dick Jeffrey (and possibly meant by Dan Garber when he talks about the Bayesian hand-held calculator). In a given context I have a number of full beliefs which delimit the presently contemplated range of possibilities; it is the latter which I grade with respect to their comparative likelihood. The context may be anchored to a problem or type of problem, for which I go to this trouble. Some of the beliefs will indeed be "deepseated", and to some I subscribe so strongly that they would survive most any change of context. They are part of what I fall back on especially if I try to specify the context in which I am presently operating – for this involves seeing myself in a "larger" perspective.

A4. Infinitesimals?[10]

There is another solution on offer for most problems which two-place probability solves. That is to stick with one-place probability, but introduce infinitesimals. Any non-self-contradictory proposition can then receive a non-zero probability, though often it is infinitesimal (greater than zero but smaller than any rational fraction).

The infinitesimals solution is to say that all the non-self-contradictory propositions (that are not contrary to my full beliefs) receive not zero probability but an infinitesimal number as probability [in a non-standard model]. There is an important result, due to Vann McGee (1994) which shows that every finitely additive 2-place probability function $P(A \mid B)$ is the standard part of $p(A \cap B)/p(B)$ for some non-standard 1-place probability function p (and conversely). Despite this I see advantages to the present approach to conditional probability which eschews infinitesimals. First of all, there is really no such thing as "the" infinitesimals

[10]For related critiques of the 'infinitesimals' gambit, see Skyrms (1983), Hajek (1992).

solution. If you go to non-standard models, there will in principle be many ways to generalize the old or ordinary concept of measure so as to keep agreement with measure in the embedded standard models. You don't have a specific solution till you specify exactly which reconstruction you prefer, and what its properties are.

In addition, in the present approach it is very easy to see how you can have "layering" of the following sort. Task: produce a two-place probability measure P such that for given A, B, C, the following is the case:

$$P(A) = 1, P(B|A) = 0, P(C|A) = 0$$
$$A,\ B,\ C \text{ are normal}$$
$$\text{If } P(X|C) = P(X|B) = P(X|A) = 0 \text{ then } X \text{ is abnormal}$$

It is easy to construct a small lexicographic model in which this is the case. Let C be a subset of B and B a subset of A; let p_1 give 1 to A but 0 to B and to C; p_2 give 1 to B but 0 to C; and p_3 give 1 to C. If these are all the measures in the sequence, then subsets of C which receive probability 0 conditional on C are all abnormal. Intuitively it would seem that in the infinitesimal approach this would require the construction in which there are exactly two layers L and M of infinitesimals: x in L is infinitesimal in comparison to standard numbers, Y in M is infinitesimal in comparison (even) to any number in L, and no numbers at all are infinitesimal in comparison to numbers in M. I leave this as an exercise for the reader.

As to the problem of belief, I wonder if the nonstandard reconstruction would have the desirable features for which we naturally turn to infinitesimals. Suppose for example that I choose a model in which each non-empty set has a positive (possibly infinitesimal) probability. Then my full beliefs are not just those which have probability 1, since that includes the tautology only. On the other hand, I can't make it a requirement that my full beliefs have probability $> 1 - d$, for any infinitesimal d one could choose. For the intersection of the sets with probability $\geq 1 - d$ will generally have a lower probability. Hence the lottery paradox comes back to haunt us. We would again face the trilemma of either restricting full beliefs to the tautology, or specifying them in terms of some factor foreign to the degrees-of-belief framework, or banishing them from epistemology altogether.

References

Birkhoff, G. (1967). *Lattice theory* (3rd ed.). Providence: American Mathematical Society.

Blume, L., Brandenburger, A., & Deckel, E. (1991a). Lexicographic probabilities and choice under uncertainty. *Econometrica, 59*, 61–79.

Blume, L., Brandenburger, A., & Deckel, E. (1991b). Lexicographic probabilities and equilibrium refinements. *Econometrica, 59*, 81–98.

De Finetti, B. (1936). Les probabilités nulles. *Bulletin des sciences mathématiques, 60*, 275–288.

De Finetti, B. (1972). *Theory of probability* (2 vols.). New York: Wiley.

Field, H. (1977). Logic, meaning and conceptual role. *Journal of Philosophy, 74*, 374–409.

Foley, R. (1993). *Working without a net*. Oxford: Oxford University Press.

Gardenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge: MIT Press.

Hajek, A. (1992). *The conditional construal of conditional probability*. Ph.D. dissertation, Princeton University.

Harper, W. L. (1976). Rational belief change, Popper functions, and counter-factuals. In C. Hooker, & W. Harper (Eds.), *Foundations of probability theory* (Vol. 1, pp. 73–112). Dordrecht: Reidel Publishing Company.

Kappos, D. A. (1969). *Probability algebras and stochastic spaces*. New York: Academic Press.

Maher, P. (1990). Acceptance without belief. *PSA, 1*, 381–392.

McGee, V. (1994). Learning the impossible, In Eells, E. & Skyrms, B. (Eds.). *Probabilities and conditionals: Belief revision and rational decision* (pp. 179–199). Cambridge: Cambridge University Press.

Popper, K. (1959). *The logic of scientific discovery*. New York: Basic Books.

Renyi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Hungarica, 6*, 285–333.

Renyi, A. (1970a). *Foundation of probability*. San Francisco: Holden-Day.

Renyi, A. (1970b). *Probability theory*. Amsterdam: North-Holland.

Skyrms, B. (1983). Three ways to give a probability function a memory. In J. Earman (Ed.), *Testing scientific theories* (Minnesota studies in the philosophy of science, Vol. X, pp. 157–161). Minneapolis: University of Minnesota Press.

van Fraassen, B. C. (1979). Foundations of probability: A modal frequency interpretation. In G. Toraldo di Francia (Ed.), *Problems in the foundations of physics* (pp. 344–387). Amsterdam/New York: North-Holland Publishing Company.

van Fraassen, B. C. (1981a). Probabilistic semantics objectified: I. Postulates and logics. *Journal of Philosophical Logic, 10*, 371–394.

van Fraassen, B. C. (1981b). Probabilistic semantics objectified: II. Implications in probabilistic model sets. *Journal of Philosophical Logic, 10*, 495–510.

van Fraassen, B. C. (1988). Identity in intensional logic: Subjective semantics. In U. Eco et al. (Eds.), *Meaning and mental representation* (pp. 201–219). Bloomington: Indiana University Press.

Vickers, J. M. (1988). *Chance and structure*. Oxford: Oxford University Press.

Chapter 6
A Theory of Higher Order Probabilities

Haim Gaifman

Introduction

The assignment of probabilities is the most established way of measuring uncertainties on a quantitative scale. In the framework of subjective probability, the probabilities are interpreted as someone's (the agent's) degrees of belief. Since justified belief amounts to knowledge, the assignment of probabilities, in as much as it can be justified, expresses knowledge. Indeed, knowledge of probabilities, appears to be the basic kind of knowledge that is provided by the experimental sciences today.

This is knowledge of a partial, or incomplete, nature, but not in the usual sense of "partial". Usually we mean by "partial knowledge" knowledge of some, but not all, of the facts in a certain domain. But knowing that a given coin is unbiased does not enable one to deduce any non-tautological Boolean combination of propositions which describe outcomes in the next, say 50 tosses. And yet it constitutes very valuable knowledge about these very same outcomes. What is the objective content of this knowledge? What kind of fact is the fact that the true probability of "heads" is 0.5, i.e., that the coin is unbiased? I shall not enter here into these classical problems[1]. I take it for granted that, among probability assignments, some are more successful, or better tuned to the actual world, than

A part of this paper has been included in a talk given in a NSF symposium on foundations of probability and causality, organized by W. Harper and B. Skyrms at UC Irvine, July 1985. I wish to thank the organizers for the opportunity to discuss and clarify some of these ideas.

[1]My Salzburg paper (1983) has been devoted to these questions. The upshot of the analysis there has been that even a "purely subjective" probability implies a kind of factual claim, for one can

H. Gaifman (✉)
Columbia University, New York, NY, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_6

others. Consequently probability assignments are themselves subject to judgement and evaluation. Having, for example, to estimate the possibility of rain I might give it, going by the sky's appearance, 70 %. But I shall be highly uncertain about my estimate and will adopt the different value given, five minutes later, in the weather forecast.

Thus we have two levels of uncertainty:

1. Uncertainty concerning the occurrence of a certain event – expressed through the assignment of probabilities.
2. Uncertainty concerning the probability values assigned in 1.

When this second level is itself expressed by assigning probabilities we get second order probabilities. An example of a second order probability is furnished by a cartoon in "The New Yorker" showing a forecaster making the following announcement:

> There is now 60 % chance of rain tomorrow, but, there is 70 % chance that later this evening the chance of rain tomorrow will be 80 %.

Just as we can iterate modal or epistemic operators, so in the system to be presented here we can iterate the probability-assignment operator to any depth. The goal of this paper is to present a general and adequate semantics for higher order probabilities and to obtain, via representation theorems, nice easily understood structures which give us a handle on the situation.

The basic structure to be defined here is a **HOP** (Higher Order Probability space). A *simple* HOP is based on a field of events, F, and on a binary operator $PR(,)$ which associates with every event A and every real closed interval Δ an event $PR(A,\Delta)$ in F. The intended meaning is that $PR(A,\Delta)$ is the event that A's true probability lies in the interval Δ.

"True probability" can be understood here as the probability assigned by an ideal expert or by someone fully informed. It is however up to us (or to the agent) to decide what in the given context constitutes an "ideal expert" or "someone fully informed". If "full information" means knowing *all* the facts then, of course, the true (unknown to us) probability has only two values 0 and 1; this will make the HOP trivial in a certain sense. In the other extreme, the agent may regard himself as being already fully informed and this leads to the "opposite" trivialization of the HOP. Generally, the agent will regard the expert as being more knowledgeable than himself, but not omniscient; e.g., the expert might know the true bias of a coin but not the outcomes of future tossings, or he might have statistical information for estimating the bias, which the agent lacks.

The agent himself at some future time can be cast in the role of being "fully informed". Thus, if P is the forecaster's present probability function and if PR

asses its success in the actual world. Rather than two different kinds, subjective and objective probabilities are better to be regarded as two extremes of a spectrum.

represents his state of knowledge later in the evening, then his announcement in "The New Yorker" cartoon can be summed up as follows, where $A =$ 'tomorrow it will rain':

$$P(A) = .6 \;\; P(PR(A, [.8, .8])) = .7$$

In order to represent knowledge at different stages, we make PR into a 3-place operator: $PR(A, t, \Delta)$ is the event that the probability of A at stage t lies in Δ. The stages can be time-points, in which case t ranges over some ordered set. More generally, the set of stages is only partially ordered, where $s \leq t$ if the knowledge at stage t includes the knowledge at stage s. (Different agents may thus be represented in the structure.) This is how a HOP is defined in general. We shall first establish the properties of simple HOPs, then use them to derive those of the more general spaces.

We shall also define, in a separate section, a formal logical calculus, to be called *probability logic*, which is naturally associated with simple HOPs. Various modalities can be reconstructed within this calculus. The general HOPs give rise to stage-dependent modalities whose calculus will be outlined at the end of the paper.

The import of the subject for various branches of philosophy and for the foundations of probability is obvious. Also obvious should be its bearing upon applied probabilistic reasoning in distributed networks, or upon efforts to incorporate such reasoning in AI systems. Mathematically, most of this paper is rather easy. Our goal has not been to prove difficult theorems, but to clarify some basic concepts and to outline a general, conceptually "clean", framework within which one can use freely and to good effect statements such as: 'With probability 0.7 Adam will know at stage 3 Bob's probability for the event A, with error ≤ 0.01' (where Adam and Bob are either people or processors). Statements of this form express intuitive thinking which may underly involved technical proofs; to use them openly and precisely can help us as a guide for finding and organizing our arguments.

A theoretic framework for higher order probabilities may also yield insights into systems of reasoning which employ non-probabilistic certainty measures. For when probability is itself treated like a random variable, we can use various methods of "safe" estimation which do not necessarily yield a probability measure. For example, define the *certainty measure* of an event A to be the largest α such that, with probability 1, the probability of A is $\geq \alpha$. This is only one, apparently the most conservative, measure among various measures that can be used.

Higher order probabilities have been considered by De-Finetti, but rejected by him owing to his extreme subjectivist views. Savage considered the possibility but did not take it up, fearing that the higher order probabilities will reflect back on the ground level, leading to inconsistencies. Instances of higher order probabilities figure in works of Good (1965) and Jaynes (1958). More recent philosophical works are by Domotor (1981), Gardenfors (1975) (for qualitative probabilities), Miller (1966), Skyrms (1980a, b) – who did much to clarify matters, van-Frassen (1984), and others.

Due to limitations of space and deadline I have not entered into details of various proofs. Some of the material has been abridged; I have included some illustrative examples of simple HOPs, but not the more interesting ones of general HOPs (which arise naturally in distributed systems). Also the bibliography is far from complete.

Simple HOPs

Definition and Basic Properties

As in Kolmogoroff's framework (1933) we interpret propositions as subsets of some universal set, say W, and we refer to them as events. We can regard W as the set of all possible worlds. Thus we have $X =$ set of all worlds in which X is true and we get the following standard correspondence:

$\vee$ (disjunction) $\mapsto \cup$ (union)
$\wedge$ (conjunction) $\mapsto \cap$ (intersection)
$\neg$ (negation) $\mapsto -$ (complementation)

Terminology A Boolean Algebra (of sets) is a class of sets closed under finite unions (and intersections) and under complementation (with respect to some presupposed universal set, in our case – W). A field is a Boolean algebra closed under countable unions (what is known also as a σ-algebra). The field (Boolean algebra) generated by a class S of sets is the smallest field (Boolean algebra) which contains S as a subclass. Note that in generating a Boolean algebra we apply finitary operations only, whereas in generating a field infinitary countable operations are used. A field is countably generated if it has a countable set of generators. All probabilities are assumed here to be countably additive.

A **HOP** is a 4-tuple (W, F, P, PR), where F is a field of subsets of W, to be called events, P is a probability over F and PR is a mapping associating with every $A \in F$ and every real closed interval Δ an event $PR(A, \Delta)$,

$$PR : F \times set\ \ of\ \ closed\ \ intervals \rightarrow F$$

As explained in the introduction $PR(A, \Delta)$ is the event that the true (or the eventual, or the expert-assigned) probability of A lies in Δ. P is the agent's current subjective probability.

Among the closed intervals, we include also the empty interval, $\varnothing$. The minimal and maximal elements of F are, respectively, $\mathbf{0}$ and $\mathbf{1}$; that is: $\mathbf{0} =$ empty subset of $W =$ False, $1 = W =$ True.

In the explanations I shall use "probability" both for the agent's current subjective probability as well as for the true, or eventual one; the contexts indicate the intended reading.

The following axioms are postulated for a HOP:

(I) $PR(A, [0,1]) = \mathbf{1}$ (For every A, the event that A's probability lies in $[0,1]$ is W, i.e., true.)

(II) $PR[A, \varnothing] = \mathbf{0}$ (That A's probability lies in the empty interval is the empty event, i.e., false.)

(III) *If* $\Delta_1 \cup \Delta_2$ is an interval then $PR(A, \Delta_1 \cup \Delta_2) = PR(A, \Delta_1) \cup PR(A, \Delta_2)$ (A's probability lies in the interval $\Delta_1 \cup \Delta_2$ iff it lies either in Δ_1 or in Δ_2)

In the following two axioms "n" is a running index ranging over $\{1,2, \ldots\}$.

(IV) $\cap_n PR(A, \Delta_n) = PR(A, \cap_n \Delta_n)$ (A's probability lies in every Δ_n iff it lies in their intersecton).

(V) *If, for all* $n \neq m$, $A_n \cap A_m = \varnothing$, *then* $\cap_n PR(A, [\alpha_n, \beta_n]) \subset PR(\cup_n A_n, [\Sigma_n \alpha_n, \Sigma_n \beta_n])$ (For pairwise disjoint A_ns, if A_n's probability lies in $[\alpha_n, \beta_n]$, $n = 1,2, \ldots,$ then the probability of $\cup_n (A_n)$ lies in $[\Sigma_n \alpha_n, \Sigma_n \beta_n]$)).)

Note that axioms (I)–(V) involve only W, F and PR. The crucial axiom which connects PR with P will be stated later.

Theorem 6.1 *For every HOP,* $H = (W, F, P, PR)$ *there is a mapping* p *which associates with every* x *in* W *a probability,* P_z, *over* F *such that*

$$PR(A, \Delta) = \{x : p_x(A) \in \Delta\} \tag{6.1}$$

The mapping p *is uniquely determined by* (6.1) *and can be defined by:*

$$p_x(A) = \inf \{\alpha : x \in PR(A, [0, \alpha])\} \tag{6.2}$$

as well as by:

$$p_x(A) = \sup \{\alpha : x \in PR(A, [\alpha, 1])\}. \tag{6.2'}$$

Vice versa, if, for every $x \in W$, P_x *is a probability over* F *such that* $\{x : p_x(A) \in \Delta\}$ *is in* F *for all* $A \in F$ *and all real closed,* Δ, *and if we use* (6.1) *as a definition of* PR *then Axioms* (I)–(V) *are satisfied.*

We call p the kernel of the HOP.

The proof of Theorem 6.1 is nothing more than a straight-forward derivation of all the required details from the axioms, using (6.2) as the definition of p_x. (The "vice versa" part is even more immediate than the first part.)

We can now extend PR and define $PR(A, \Xi)$, for arbitrary subsets Ξ of reals, as $\{x : p_z(A) \in \Xi\}$. If Ξ is a Borel set then $PR(A, \Xi)$ is in F.

The meaning of p_x is obvious: it is the probability which corresponds to the maximal state of knowledge in world x – the distribution chosen by the expert of that world.

Notation For $\alpha \in [0,1]$, $PR(A, \alpha) =_{\text{df}} PR(A, [\alpha, \alpha])$.

The picture is considerably simpler in the discrete case, where W is countable. Assuming with no loss of generality that $\{x\} \in F$ for every $x \in W$, the probability of some $A \subset W$ is simply the sum of the probabilities of the worlds in A. In that case, we can eliminate the closed intervals and consider only the special cases $PR(A,\alpha)$ where α ranges over $[0,1]$; also our 5 axioms can be replaced by 3 simpler ones. Discrete cases arise in many situations and are very useful as illustrative examples. But to consider only the discrete case is highly restrictive.

Notation For $x,y \in W$, $A \in F$, put: $p(x,A) =_{\text{df}} p_x(A)$ and (assuming $\{y\} \in F$) $p(x,y) =_{\text{df}} p(x,\{y\})$ and $P(y) =_{\text{df}} P(\{y\})$.

In the discrete case P is obviously determined by the values $P(x)$, $x \in W$. Thus, ordering W, we can represent P as a probability vector (a countable vector of non-negative entries which sum up to 1). Similarly the kernel p becomes a probability matrix (a countable square matrix in which every row is a probability vector). Examples (i) and (ii) in the Examples subsection can serve to illustrate the situation (the discussion there presupposes however the next subsection).

Mathematically, what we have got is a Markov process (with initial probability P and transition probabilities $p(x,)$, $x \in W$). But the interpretation is altogether different from the usual interpretation of such a structure. The connection between P and the kernel p is established in the sixth axiom.

Axiom (VI) and Its Consequences

Let $P(A|B)$ be the conditional probability of A, given B. It is defined in the case that $P(B) \neq 0$ as $P(A \cap B)/P(B)$. It is what the agent's probability for A should be had he known B.

Axiom (VI_w) *If* $P(PR(A, [\alpha,\beta])) \neq 0$ *then* $\alpha \leq P(A \mid PR(A, [\alpha, \beta])) \leq \beta$.

Axiom (VI_w) (the weak form of the forthcoming Axiom (VI)) is a generalization of Miller's Principle to the case of interval-based events. Rewritten in our notation, Miller's Principle is: $P(A \mid PR(A,\alpha)) = \alpha$. Axiom ($\text{VI}_\text{w}$) appears to be the following rule: My probability for A should be no less than α and no more than β, were I to know that in a more informed state my probability for A will be within these bounds. Plausible as it sounds, the use of the hypothetical "were I to know that . . . " needs in this context some clarification. Now a well-known way of explicating conditional probabilities is through conditional bets. Using such bets van-Frassen (1984) gives a Dutch-book argument for the Principle: Its violation makes possible a system of bets (with odds in accordance with the agent's probabilities) in which the agent will incur a net loss in all circumstances. In this argument $PR(A,\alpha)$ is interpreted as the event that the agent's probability for A at a certain future time will be α, in which case he should accept at that time bets with odds α. The same kind of Dutch-book can be constructed if Axiom (VI_w) is violated. (Here it is crucial that we use an interval, the argument fails if we replace $[\alpha,\beta]$ by a non-convex Borel set.)

Axiom (VI) is the interval-based form of the stronger version of Miller's Principle which was suggested by Skyrms (1980a).

Axiom (VI) *If C is a finite intersection of events of the form PR(B,Δ), and if P(C ∩ PR(A,[α,β])) ≠ 0, then*

$$\alpha \leq P\left(A \middle| C \cap PR\left(A,[\alpha,\beta]\right)\right) \leq \beta$$

The same intuition which prescribes (VI_w) prescribes (VI); also here the violation of the axiom makes possible a Dutch-book against the agent. What is essential is that events of the form $PR(B,\Delta)$ be, in principle, knowable to the agent, i.e., be known (if true) in the maximal states of knowledge as defined by our structure.[2]

In what follows integrating a function $f(t)$ with respect to a probability m is written as $\int \mathrm{f}(t) \cdot m(dt)$.

Lemma 6.1 *Axiom* (VI_w) *implies that the following holds for all A ∈ F:*

$$P(A) = \int p\,(x,A) \cdot P(dx) \tag{6.3}$$

The proof consists in applying the formula $P(A) = \Sigma_i P(A \mid B_i) \cdot P\,(B_i)$, where the B_i's form a partition, passing to the limit and using the definition of an integral.

The implication (6.3) ⇒ (VI_w) is not true in general. Note that in the discrete case (6.3) becomes:

$$P(x) = \sum\nolimits_y p\,(x,y) \cdot P(y) \tag{6.3_d}$$

which means that the probability vector is an eigen-vector of the kernel.

Definition Call Two worlds $x,y \in W$ epistimically equivalent, (or, for short, equivalent) and denote it by $x \simeq y$, if $P_x = P_y$ • For S – a class of events, define $K[S]$ to be the field generated by all events of the form $PR(A,\Delta)$, $A \in S$, Δ – a real closed interval.

Epistemic equivalence means having the same maximal knowledge. Evidently $x \simeq y$ iff, for all A and all Δ, $x \in \mathrm{PR(A,\Delta)} \Leftrightarrow y \in \mathrm{PR(A,\Delta)}$. This is equivalent to: for all $C \in K[F]$, $x \in C \Leftrightarrow y \in C$. If $K[F]$ is generated by the countably many generators X_n, $n = 0,1,\ldots$ then the equivalence classes are exactly all non-empty intersections $\cap_n X_n$' where each X_n' is either X_n or its complement. Hence the equivalence classes are themselves in $K[F]$, they are exactly the atoms of this field. The next lemma

[2]It is important to restrict C in Axiom (VI) to an intersection of such events. The removal of this restriction will cause the p_x's to be two-valued functions, meaning that all facts are known in the maximal knowledge states.

shows that the condition that $K[F]$ be countably generated is rather mild, for it holds whenever F itself is countably generated (which is the common state of affairs):

Lemma 6.2 *If S is either countable or a countably generated field, then $K[S]$ is countably generated.*

(As generators for $K[S]$ one can take all $PR(A,\Delta)$, $A\in S$, Δ – a rational closed interval; the second claim is proved by showing that if S' is a Boolean algebra that generates the field S then $K[S'] = K[S]$.)

Terminology A 0-set is a set of probability 0. Something is said to hold for almost all x if it holds for all x except for a 0-set. The probability in question is P, unless specified otherwise.

Theorem 6.2 *If F is countably generated then axiom* (VI) *is equivalent to each of the following conditions:*

(A) (6.3) *holds* (*for all A*) *and the following is true: Let C_x be the epistemic equivalence class to which x belongs, then*

$$p_x(C_x) = 1 \text{ for almost all } x.$$

(B) (6.3) *holds and, for almost all x, for all A:*

$$p_x(A) = \int p_y(A)\cdot p_x(dy) \tag{6.4}$$

The proof that axiom (VI) is equivalent to (A) and implies (B) uses only basic measure theory. The present proof of (B) $\Rightarrow$ (A) relies on advanced ergodic theory[3] and I do not know if this can be avoided. Fortunately the rest of this paper does not rely on this implication (except the corresponding implication in Theorem 6.3). Note that in the discrete case (6.4) is equivalent to:

$$p(x,z) = \sum\nolimits_y p(x,y)\cdot p(y,z) \tag{4_d}$$

(4_d) means that the kernel, as a matrix, is equal to its square.

Let $\{E_u : u \in \cup\}$ be the family of epistemic equivalence classes, with different indices attached to different classes. Let P_u be the common p_x for $x \in E_u$; let m be the probability, defined for all $V \subset \cup$ such that $\mathrm{U}_{u\in U}E_u \in F$, by:

$$m(V) = P(\cup_{u\in \mathrm{V}}E_u)$$

[3]I am thankful to my colleagues at the Hebrew University H. Furstenberg, I. Katzenelson and B. Weiss for their help in this item. Needless to say that errors, if any, are my sole responsibility.

Then (A) is equivalent to the following condition:

(C) *For all A in F*

$$P(A) = \int_U P_u(A) \cdot m(du)$$

and for almost all (with respect to m) u

$$P_u(E_u) = 1$$

The first equality in (C) is a recasting of (6.3); it can be equivalently described by saying that P is a mixture of the P_u's with weight function m. Altogether (C) means that we have here what is known as the disintegration of the probability space. It makes for a rather transparent structure.

For W – countable the disintegration means the following: After deleting from the kernel-matrix rows and columns which correspond to some set of probability 0, the rest decomposes into submatrices around the main diagonal in each of which all rows are equal, with 0's in all other places; P itself is a mixture of these rows. Such HOPs are exactly those that can be constructed as follows (hence this is the method for setting up higher order probabilities which avoid a Dutch book):

- *Chose a partition $\{E_u : u \in U\}$ of W into non-empty disjoint sets, with different u's marking different sets.*
- *Chose for each u in U a probability, $P_{u'}$ on W such that $P_u(E_u) = 1$ for all $u \in U'$, where U' is some non-empty susbset of U.*
- *Chose a probability, m, on U such that $m(U') = 1$, and let P be the mixture of the P_u's with weight function m.*
- *For each $u \in U$ and each $x \in E_x$ put $p_x = P_x$ and define $PR(A, \Delta)$ to be $\{x : p_x(A) \in \Delta\}$.*

The construction is essentially the same for a general W (with a countably generated F); some additional stipulations of measurability should be included in order to make possible the formation of the mixture and to ensure that the $PR(A, \Delta)$'s are in F.

Definition Call P_u and its corresponding equivalence class, E_u, ontological if $P_u(E_u) = 1$, call it and its corresponding class coherent if P_u is a mixture of ontological P_u's. Call a world ontological (coherent) if it belongs to an ontological (coherent) equivalence class.

An ontological class is of course coherent. A coherent class which is not ontological must get the value 0 under its own P_u. It represents a state of knowledge in which the agent knows for sure that his eventual probability function will be different from his current one (and that it will be an ontological one).

The set of ontological worlds gets the value 1 under P and under each p_x where x is coherent. It is referred to as the ontological part of the HOP. Together with the structure induced by the original HOP it forms by itself a simple HOP. Similarly we define the coherent part of the HOP as the set of all coherent worlds (together with the induced structure). As far as calculating probabilities goes, only the ontological part matters. Coherent non-ontological worlds are useful as representatives of transitory states of knowledge.

Examples

Example 6.1 $W = \{w1, w2, w3\}$ $P = (1/3, 1/3, 1/3)$ and the kernel matrix is:

$$\begin{matrix} .5 & .5 & 0 \\ 0 & .5 & .5 \\ .5 & 0 & .5 \end{matrix}$$

The agent's current probability assigns each world the value 1/3. Eventually, in world *w1* he will know that he is not in *w3* and he will assign each of the worlds *w1, w2* the value 0.5. This is the meaning of the first row. The other rows are similarly interpreted.

By direct checking one can verify that (VI_w) is satisfied. (The checking of all cases in this example is easy because $PR(A, \alpha) \neq \varnothing$ only for $\alpha = 0.5,1.$) However the matrix is not equal to its square, hence Axiom (VI) is violated, as indeed the following case shows: Put $A = \{w1\}$, $C = PR(\{w2\},0.5)$. Then $C = \{x : p(x, w2) = 0.5\} = \{w1, w2\}$ and similarly $PR(A,0.5) = \{w1, w3\}$. Hence $A = PR(A,0.5) \cap C$ implying $P(A \mid PR(A,0.5) \cap C) = 1 \neq 0.5$. This can be used to construct a Dutch book against the agent.

Note also that the epistemic equivalence classes are $\{w1\}$,$\{w2\}$ and $\{w3\}$ and that non is ontological; hence also there are no coherent worlds here.

Example 6.2 $W = \{w1,w2, \ldots,w8\}$, P is: (.1, .2, .2, .1, .4, 0, 0, 0} and the kernel matrix is:

$$\begin{matrix} .2 & .4 & .4 & & & & & \\ .2 & .4 & .4 & & & & & \\ .2 & .4 & .4 & & & & & \\ & & & .2 & .8 & & & \\ & & & .2 & .8 & & & \\ & & & & & 1 & & \\ .05 & .2 & .2 & .05 & .2 & .5 & 0 & 0 \\ .2 & .1 & .1 & .2 & .1 & .1 & .1 & .1 \end{matrix}$$

where all undisplayed entries are 0. The sets $\{w1, w2, w3\}$, $\{w4,w5\}$ and $\{w6\}$ are equivalence classes which are ontological. P is a mixture of these 3 types of rows,

with weights 0.5, 0.5, 0, respectively. Hence condition (C) is satisfied, therefore also Axiom (VI). *w7* is a coherent non-ontological world, because the 7th row is a mixture of the first three types (with weights .25, .25, .5) *w8* is not coherent. The ontological part consists of the upper left 6×6 matrix and the coherent part of the 7×7 one.

The example can be made more concrete by the following scenario. A number is to be chosen from $\{1,2,3\}$. For $i = 1,2,3$, the number chosen in wi is i, but in each of these 3 worlds the maximal knowledge consists in assigning probabilities 0.2, 0.4, 0.4 to the 3 possibilities. In *w4* the number chosen is *1* and in *w5* it is 2; in either of these worlds the maximal knowledge consists in assigning the probabilities 0.2, 0.8. In *w6* the number is 2 and it is also assigned probability 1. In the agent's current state he assigns probability 0 to finding himself eventually in the third state of maximal knowledge, and equal probabilities to the first and second states. World *w7* represent a similar situation but with different weights. We can imagine 3 lotteries for chosing the number; in each equivalence class the maximal knowledge is knowledge of the chosen lottery.

Example 6.3 Let H be the probability of "heads" of some given coin of unknown bias. Treat H as a random variable. The agent's knowledge is represented by a probability distribution for H. Say it is the uniform distribution over [0,1]. The expert does not know the value of H but he has some additional information. Say his additional information is the value of N – the number of "heads" in 50 independent tosses. Then our worlds can be regarded as pairs (h,n), such that in (h,n) the event $H = h \cap N = n$ is true; here h is a real number in [0,1] and n an integer between 0 and 50. The field F is generated by the sets $[\alpha, \beta] \times \{n\}, 0 \leq \alpha \leq \beta \leq 1, \ n = 0, \ldots, 50$.

Given $H = h$, we get the binomial distribution $b_{h,50}$ for N. This fact, together with the agent's uniform distribution for H, determines his probability P over F. The expert's probability in world (h,n) is obtained by conditioning on his information, it is $P(\ |N = n)$. There are 51 equivalence classes which correspond to the 51 possible values of N and all worlds are ontological.

As is well known, different values of N give rise to different conditional distributions of H. Therefore the events $N = n$ are in the field generated by the events[4] $PR(H \in [\alpha,\beta], \ \Delta)$. The whole field F is therefore generated by events which are either of the form $H \in [\alpha,\beta]$ or obtained from these by applying the operator PR. Consequently we can give an abstract description of this HOP which does not mention the fifty tosses. The only function of the tosses is to affect the distribution of H; in our framework such changes in distribution constitute themselves events which can be treated directly, without having to bring in their causes.

[4] Actually there are 51 real numbers α_n such that the event $N = n$ is the same as $PR(H \leq 1/2, \alpha_n)$.

The Case of a General Field

The restriction that F be countably generated is a mild one. The probability spaces which commonly appear in theory, or in applications, are essentially of this nature[5]. Usually we are interested in properties that involve only countably many generators. We will first show that for studying such properties we can always restrict ourselves to the case where the underlying field is countably generated.

Definition Given a simple HOP (*W, F, P, PR*) and given $S \subset F$, define $H[S]$ as the smallest field containing S and closed under PR (i.e., $A \in H[S] \Rightarrow PR(A, \Delta) \in H[S]$ for every real closed interval Δ).

$H[S]$, together with the restrictions of P and PR to it, forms a subHOP, where this notion is defined in the obvious way.

Lemma 6.3 *If S is a Boolean algebra and, for every A in S and every rational closed interval* Δ, $PR(A, \Delta)$ *is in S, then* $H[S]$ *is the field generated by S.*

This means that, once we have a Boolean algebra closed under $PR(\ ,\Delta)$ for all Δ with rational endpoints, we get all the rest by countable Boolean operations without using *PR*.

Corollary *If S is either countable, or a countably generated field, then* $H[S]$ *is countably generated.*

Using this we can derive from Theorem 6.2 an analogous result for general fields:

Theorem 6.3 *Axiom* (VI) *is equivalent to each of the following conditions:*

(A′) (3) *holds and for every C in* $K[F]$, *for almost all x:* $p_x(C) = 1$ *if* $x \in C$, $p_x(C) = 0$ *otherwise.*
(B′) (3) *holds and for every A in F* (4) *is true for almost all x.*
(B′) differs from the analogous (B) of Theorem 6.2 in that the exceptional 0-set for (4) can depend on *A*

Say that *A is equal a.e. to B* if *A-B* and *B-A* are 0-sets. Say that two classes of sets are *equal modulo 0-sets* if every member of one is equal a.e. to some member of the other.

Assuming Axiom (VI) we get:

Corollary *If* $S \subset F$, *then: (i) The fields* $K[S]$, $K[K[S]]$ *and* $K[H[S]]$ *are equal modulo 0-sets. (ii) If S is a boolean algebra then* $H[S]$ *is equal modulo 0-sets to the field generated by* $S \cup K[S]$.

(To show, for example, that $K[S] = K[K[S]]$ modulo 0-sets, consider $C \in K[S]$; by Theorem 6.3, $\{x : p_x(C) \in \Delta\}$ is equal a.e. to C if $\Delta = [1,1]$, is equal a.e. to $W\text{-}C$ if

[5]They are *seperable,* i.e., for some countably generated field every event in the space differs from a set in the field by a 0-set.

$\Delta = [0,0]$, and is a 0-set if $0,1 \notin \Delta$. Hence, for all Δ, $PR(C,\Delta)$ is equal a.e. to one of: C, W-C, W, $\varnothing$. Since $K[K[S]]$ is generated by such sets, the claim follows.)

Roughly speaking, (ii) means that, modulo 0-sets, nested applications of PR reduce to non-nested applications. A stronger, syntactical version of this is given in the next section.

Probability Logic

Let Ξ be a set of reals such that $0,1 \in \Xi$. Call an interval with end-points in Ξ a Ξ-*interval*.

Let PRL_Ξ be the calculus obtained by adjoining sentential operants, $PR(\ ,\Delta)$, to the propositional calculus, where Δ ranges over all closed Ξ-intervals. Here, for the sake of convenience, I use 'PR' for the syntactical operant, as well as for the operation in HOPs. Given some class $\{X_i\colon i \in I\}$ of sentential variables, the class of all wffs (well formed formulas) of PRL_Ξ is the smallest such that:

- Every sentential variable is a wff
- If ϕ and ψ are wffs, so are $\neg\phi$ and $\phi * \psi$ where * is any of the standard binary connectives.
- If ϕ is a wff and Δ is a closed Ξ-interval then $PR(\phi,\Delta)$ is a wff.

Let $\mathrm{H} = (W,F,P, PR)$ be a simple HOP and let r be a mapping which maps each sentential variable to a member of F. Then the value $|\phi|_{\mathrm{H},r}$ of the wff ϕ is defined by interpreting the sentential connectives as the corresponding Boolean operations and each syntactic operant $PR(\ ,\Delta)$ as the operation $PR(\ ,\Delta)$ of the HOP.

Definition A wff ϕ, is *p-valid*, to be denoted $\models_p \phi$, if, for every simple HOP H which satisfies Axiom (VI) and every r, the probability of $|\phi|_{\mathrm{H},\tau}$ is 1. Two wffs ϕ, ψ are *p-equivalent* if $\phi \leftrightarrow \psi$ is p-valid.

Call ϕ a *PC-formula* if it is a wff of the propositional calculus, i.e., does not contain any PR.

Theorem 6.4 *Every wff of* PRL_Ξ *is p-equivalent to a Boolean combination of PC-formulas and formulas of the form* $PR(\sigma,\Delta)$ *in which* σ *ranges over PC-formulas.*

This means that as far as probabilities are concerned (i.e., if we disregard 0-sets) we need not use nested PR's.

Theorem 6.5 *Translate into* PRL_Ξ *the wffs of propositional modal logic with the necessity oprant N, by replacing each* $N(\psi)$ *by* $PR(\psi, [1,1])$. *Let* ϕ^* *be the translation of* ϕ. *Then*

$$S5 \vdash \phi \text{ iff } \models_p \phi^*$$

Thus S5 becomes a fragment of PRL_Ξ. This relation becomes more explicit if we rewrite '$PR(\psi,\Delta)$' as '$N_\Delta(\psi)$'.

It can be shown that for $\Xi =$ set of rationals the set of p-valid wffs is recursive. Also PRL_{Ξ} can be provided with a natural set of formal axioms so that, with modus ponens as derivation rule, p-validity coincides with provability.

Some Questions
Other validity notions can be considered (e.g., that $|\phi|_{\mathrm{H},\tau}$ always contains all coherent worlds in the HOP), as well as other interpretations of the necessity operant (e.g., as $\phi \bigwedge PR(\phi,[1,1])$). What modal logics are thereby obtained?

General HOPs

In general, a HOP is a structure of the form:

$$(W, F, P, T, PR)$$

where, as before, (W F, P) is a probability space, $T = (T, <)$ is a partially ordered set and where

$$PR : F \times T \times \textit{set of closed intervals} \rightarrow F$$

$PR(A,\ t,\ \Delta)$ is the event that the probability of A at stage t lies in Δ. If the stages coincide with time points then the partial ordering of T is total. As before, P is the current subjective probability; here "current" is earlier (i.e., less than or equally informative) than the stages in T. Put:

$$PR_t\,(A, \Delta) =_{\mathrm{df}} PR\,(A, t, \Delta)$$

The first five axioms **(I*)–(V*)** in this setting are the obvious generalizations of our previous axioms (I)–(V). Namely, we replace 'PR' by 'PR_t' and require that the condition hold for all t in T.

Theorem 6.1 generalizes in the obvious way and we get, for each $t \in T$ and each $x \in W$, a probability $P_{t,x}$ which determines PR_t; it represents the maximal state of knowledge at stage t in world x.

The "correct" generalization of Axiom (VI) is not as obvious, but is not difficult to find:

Axiom (VI*) *For each $t \in T$ the following holds: If C is a finite intersection of events of the form $PR_s(B,\Delta)$ where every s is $\leq t$, and $P\,(C \cap PR_t\,(A, [\alpha, \beta])) \neq 0$, then*

$$\alpha \leq P\left(A \middle| C \cap PR_t\,(A\,[\alpha,\ \ \beta])\right) \leq \beta$$

The argument for this axiom is the same as the argument for Axiom (Vl). The essential point is that if $s \leq t$ then true events of the form $PR_t(\mathrm{B},\Delta)$ are known at stage t. The same Dutch book argument works for Axiom (VI*).

As before, we consider fields generated by knowable events and define epistemic equivalence; but now these concepts depend on the stage parameter, to be displayed here as an additional subscript. Thus we put:

$$x \simeq_t y \iff_{df} P_{t,x} = P_{t,y}$$

Then $x \simeq_t y$ iff $x \in A \Leftrightarrow y \in A$, for all $A \in K_t[F]$.

Theorem 6.6 *Assume F to be countably generated, then Axiom* (VI*) *is equivalent to the conjunction of:*

(D) *For each $t \in T$ the simple HOP (W, F, P, PR_t) satisfies Axiom* (Vl).
and
(E) *For each $s \leq t$, $x \simeq_t y \Rightarrow x \simeq_s y$, for almost all x,y (i.e., for all $x,y \in W'$ where $P(W') = 1$).*

(E) means that, as we pass to more progressive stages, almost everywhere epistemic equivalence is the same or becomes stronger; the partition into equivalence classes can change only by becoming more refined.

Like Theorem 6.2 the last theorem has a version that applies to general fields but I shall not enter here into it. In the following theorem F is assumed to be countably generated.

Theorem 6.7 *Assume Axiom* (VI*). *Let $s \leq t$. Then, for almost all x, $p_{s,x}$ is a mixture of $p_{t,y}$'s (where y ranges over W). Consequently, for almost all x, $(W, F, p_{s,x'} PR_t)$ is a simple HOP satisfying Axiom* (VI).

Logic of HOPs and Stage Dependent Modalities

Fix a partially ordered set $T = (T, <)$. The logic $\mathrm{PRL}_{\Xi,T}$ (which corresponds to HOPs with set of stages T) is defined in the same way as PRL_{Ξ}, except that PR has an additional argument ranging over T. As before we employ a systematically ambiguous notation. Define ϕ to be p-valid if it gets probability 1 in all HOPs in which the set of stages is T.

Now consider a propositional modal language, M_T, in which we have, instead of a single necessity operant, an indexed family N_t, $t \in T$. $N_t\phi$ states that ϕ is necessary at stage t, i.e., necessary by virtue of the maximal knowledge available at that stage.

For $\phi \in \mathrm{M}_T$, let ϕ^* be the wff obtained by replacing each $N_t\psi$ by $PR_t(\psi, [1,1])$. It can be shown that the set of all ϕ in M_T such that ϕ^* is p-valid is exactly the set of wffs derivable, by modus ponens and the rule: if $\vdash \phi$ then $\vdash N_t\phi$, from the following axioms:

(i) All tautologies. (ii) For each $t \in T$, the axiom schemas of S5, with N replaced by N_t and

(iii) $N_s\phi \rightarrow N_t\phi$, for each $s \leq t$.

Note that (iii) accords well with the intended meaning of the N_t's: If something is necessary at stage s it is also necessary at later stages. On the other hand, something not necessary at stage s can be necessary later.

References[6]

Domotor, Z. (1981). *Higher order probabilities* (Manuscript).

Gaifman, H. (1983). *Toward a unified concept of probability.* Manuscript of invited lecture to the 1983 International Congress for Logic Philosophy and Methodology of science, salzburg. To appear in the proceedings of the congress, North Holland.

Gardenfors, P. (1975). Qualitative probability as intentional logic. *Journal of Philosophical Logic, 4*, 171–185.

Good, I. J. (1965). *The estimation of probabilities*. Cambridge: MIT Press.

Jaynes, E. T. (1958). *Probability theory in science and engineering* (Colloquium lectures in pure and applied science No 4). Dallas: Socony Mobil Oil Co Field Research Library.

Kolmogoroff, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeit* (Ergebnisse der Mathematik 502 und ihrer Grenzgebitte, no 2). Berlin: Springer.

Miller, D. (1966). A paradox of information. *British Journal for the Philosophy of Science, 17*, 59–61.

Skyrms, B. (1980a). Higher order degree of belief. In D. H. Mellor (Ed.), *"Prospects for pragmatism" essays in honor of F.P. Ramsey*. Cambridge: Cambridge University Press.

Skyrms, B. (1980b). *Causal necessity* (*Appendix 2*). New-Haven: Yale.

van Frassen, B. (1984). Belief and the will. *Journal of Philosophy, 81*, 235–256.

[6]The list is far from being complete. Some important papers not mentioned in the abstract are to be found in Ifs, W. Harper ed. Boston Reidel, 1981. Material related in a less direct way: non-probabilistic measures of certainty (e.g., the Dempster-Shafer measure), expert systems involving reasoning with uncertainties, probabilistic protocols and distributed systems, has not been included, but should be included in a fuller exposition.

Chapter 7
On Indeterminate Probabilities

Isaac Levi

SOME men disclaim certainty about anything. I am certain that they deceive themselves. Be that as it may, only the arrogant and foolish maintain that they are certain about everything. It is appropriate, therefore, to consider how judgments of uncertainty discriminate between hypotheses with respect to grades of uncertainty, probability, belief, or credence. Discriminations of this sort are relevant to the conduct of deliberations aimed at making choices between rival policies not only in the context of games of chance, but in moral, political, economic, or scientific decision making. If agent X wishes to promote some aim or system of values, he will (*ceteris paribus*) favor a policy that guarantees him against failure over a policy that does not. Where no guarantee is to be obtained, he will (or should) favor a policy that reduces the probability of failure to the greatest degree feasible. At any rate, this is so when X is engaged in deliberate decision making (as opposed to habitual or routine choice).

Two problems suggest themselves, therefore, for philosophical consideration:

The Problem of Rational Credence: Suppose that an ideally rational agent X is committed at time t to adopting as certain a given system of sentences $K_{X,t}$ (in a suitably regimented L) and to assigning to sentences in L that are not in $K_{X,t}$ various degrees of (personal) probability, belief, or credence. The problem is to specify conditions that X's "corpus of knowledge" $K_{X,t}$ and his "credal state" $B_{X,t}$ (i.e., his system of judgments of probability or credence) should satisfy in order to be reasonable.

The Problem of Rational Choice: Given a corpus $K_{X,t}$ and a credal state $B_{X,t}$ at t, how should X make decisions between alternative policies from which he must choose one at t?

I. Levi (✉)
Columbia University, New York, NY, USA
e-mail: levi@columbia.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_7

Consideration of these two problems should lead to examination of a third. A rational agent X is entitled to count as certain at t not only logical, mathematical, and set-theoretical truths supplemented by suitably produced testimony of the senses, but theories, laws, and statistical claims as well. At the same time, the revisability of X's corpus at t should be recognized not only by others but by X himself. Moreover, just as X's judgments of certainty are liable to revision, so too are his judgments of probability or credence. Indeed, the two types of modification are apparently interdependent, and this interdependence itself deserves examination. The third problem, therefore, is as follows:

The Problem of Revision: Under what conditions should X modify his corpus $K_{X,t}$ or his credal state $B_{X,t}$, and, if he should do so, how should he choose between alternative ways of making revisions?

In this essay, I shall not attempt to solve the problem of revision. However, I shall indicate how a prima facie obstacle to offering anything other than a dogmatic or antirationalistic answer to the question can be eliminated.

The obstacle is a serious one; for it derives from a very attractive system of answers to the problem of rational credence and the problem of rational choice. I allude to what is called the "bayesian" view. Bayesians do not agree with one another in their answers to these questions in all respects. The views of Harold Jeffreys and the early views of Rudolf Carnap are not consonant in important ways with the ideas of Bruno de Finetti and Leonard J. Savage (or the later Carnap). Nonetheless, the answers these and a host of other authors offer to the first two questions share certain important ramifications for the problem of revision. One of these implications is the commitment to either dogmatism or antirationalism.

Of course, identifying an objectionable consequence of bayesianism, where the objection is grounded on a question of philosophical principle, is in itself unlikely to persuade devoted bayesians to abandon their position. Such authors will be tempted to modify philosophical principle so as to disarm the objection; and they will have good reasons for doing so. Bayesian doctrine does offer answers to the first two questions. These answers are derivable from a system of principles which are precise and simple. Even the disputes between bayesians can be formulated with considerable precision. Furthermore, the prescriptions bayesians recommend for making choices appear to conform to presystematic judgment at least in some contexts of decision. Rival attempts to answer the problems of rational credence and rational choice seem either eclectic or patently inadequate when compared with the bayesian approach.

Thus, it is not enough to complain of the defects of bayesianism. The serious challenge is to construct an alternative system of answers to the problems of rational credence and choice which preserves the virtues of bayesianism without its vices—in particular, the defects it exhibits relevant to the problem of revision.

In this paper, I shall outline just such a rival view.

I

X's corpus of knowledge $K_{X,t}$ at t identifies a set of options $A_1, A_2, \ldots, A_n$ as the options from which he will choose (at t' identical with or later than t) at least and at most one. In addition, $K_{X,t}$ implies that at least and at most one of the hypotheses h_1, $h_2, \ldots, h_m$ is true and that each of the h_j's is consistent with $K_{X,t}$. Finally, $K_{X,t}$ implies that, if X chooses A_i when h_j is true, the hypothesis o_{ij} asserting the occurrence of some "possible consequence" of A_i is true.

The problem of rational choice is to specify criteria for evaluating various choices of A_is from among those feasible for X according to what he knows at t. Such criteria may be construed as specifying conditions for "admissibility." Option A_i is admissible if and only if X is permitted as a rational agent to choose A_i from among the feasible options. If A_i is uniquely admissible, X is obliged, as a rational agent, to choose it. In general, however, unique admissibility cannot be guaranteed, and no theory of rational choice pretends to guarantee it.

Bayesians begin their answer to the problem of rational choice by assuming that X is an ideally rational agent in the following sense:

(i) X has a system of evaluations for the possible consequences (the o_{ij}s) representable by a real-valued "utility" function $u(o_{ij})$ unique up to a linear transformation (i.e., where utility assignments are nonarbitrary once a 0 point and a unit are chosen—as in the case of measuring temperature).

(ii) X has a system of assignments of degrees of credence to the o_{ij}s, given the choice of A_i representable by a real-valued function $Q(o_{ij};\ A_i)$ conforming to the requirements of the calculus of probabilities. Often X will assign credence values to the "states of nature" $h_1, h_2, \ldots, h_n$ so that the h_js are probabilistically independent of the option chosen. When this is so, $Q(o_{ij};\ A_i)$ equals the unconditional credence (given $K_{X,t}$) $Q(h_j)$. In the sequel, I shall suppose that we are dealing with situations of this kind.

Given such a utility function $u(o_{ij})$ and Q-function $Q(h_i)$, let $E(A_i) = \sum_{j=1}^{m} u\left(o_{ij}\right) Q\left(h_j\right)$. $E(A_i)$ is the expected utility of the option A_i.

Bayesians adopt as their fundamental principle of rational choice the principle that an option is admissible only if it bears maximum expected utility among all the feasible options.

Very few serious writers on the topic of rational choice object to the principle of maximizing expected utility in those cases where X's values and credal state can be represented by a utility function unique up to a linear transformation and a unique probability function. The doubts typically registered concern the applicability of this principle. That is to say, critics doubt that ordinary men have the ability under normal circumstances to satisfy the conditions of ideal rationality stipulated by strict bayesians even to a modest degree of approximation.

The bayesian riposte to doubts about applicability is to insist that rational men should meet the requirements for applying the principle of maximizing expected

utility and that, appearances to the contrary notwithstanding, men are quite capable of meeting these requirements and often do so.

I am not concerned to speculate on our capacities for meeting strict bayesian requirements for credal (and value) rationality. But even if men have, at least to a good degree of approximation, the abilities bayesians attribute to them, there are many situations where, in my opinion, rational men *ought not* to have precise utility functions and precise probability judgments. That is to say, on some occasions, we should avoid satisfying the conditions for applying the principle of maximizing expected utility even if we have the ability to satisfy them.

In this essay, reference to the question of utility will be made from time to time. I shall not, however, attempt to explain why I think it is sometimes (indeed, often) irrational to evaluate consequences by means of a utility function unique up to a linear transformation. My chief concern is to argue that rational men should sometimes avoid adopting numerically precise probability judgments.

The bayesian answer to the problem of rational choice presupposes at least part of an answer to the problem of rational credence. For a strict bayesian, a rational agent has a credal state representable by a numerically precise function on sentences (or pairs of sentences when conditional probability is considered) obeying the dictates of the calculus of probabilities.

There are, to be sure, serious disputes among bayesians concerning credal rationality. In his early writings, Carnap believed that principles of "inductive logic" could be formulated so that, given X's corpus $K_{X,t}$, X's credal state at t would be required by the principles of inductive logic to be represented by a specific Q-function that would be the same for anyone having that corpus.[1] Others (including the later Carnap[2]) despair of identifying such strong principles. Nonetheless, bayesian critics of the early Carnap's program for inductive logic continue to insist that ideally rational agents should assign precise probabilities to hypotheses.

II

X's corpus of knowledge $K_{X,t}$ shall be construed to be the set of sentences (in L) to whose certain truth X is committed at t. I am not suggesting that X is explicitly or consciously certain of the truth of every sentence in $K_{X,t}$, but only that he is committed to being certain. X might be certain at t of the truth of h and, hence, be committed to being certain of $h \vee g$, without actually being certain. Should it be brought to X's attention, however, that $h \vee g$ is a deductive consequence of h, he would be obliged as a rational agent either to cease being certain of h or to take $h \vee g$ to be certain. The latter alternative amounts to retaining his commitment; the former to abandoning it.

[1] *Logical Foundations of Probability* (Chicago: University Press, 2nd ed., 1962), pp. 219–241.

[2] "Inductive Logic and Rational Decisions," in Carnap and R. C. Jeffrey, eds., *Studies in Inductive Logic and Probability* (Berkeley: UCLA Press, 1971), p. 27.

In this sense, X's corpus of knowledge at t should be a deductively closed set of sentences. Insofar as we restrict our attention to changes in knowledge and credence which are changes in commitments, modifications of corpora of knowledge are shifts from deductively closed sets of sentences to other deductively closed sets of sentences. Such modifications come in three varieties:

1. *Expansions*, where X strengthens his corpus by adding new items. Some examples of expansion are acquiring new items via observation, from the testimony of others and through inductive or nondeductive inference leading to the "acceptance" of statistical claims, laws, or theories into the corpus.
2. *Contractions*, where X weakens his corpus by removing items. This can happen when X detects an inconsistency in his corpus due to his having added at some previous expansion step an observation report that contradicts assumptions already in his corpus, or when X finds himself in disagreement with Y (whose views he respects on the point at issue) and wishes to resolve the dispute without begging the question.
3. *Replacements*, where X shifts from a theory containing one assumption to another containing an assumption contradicting the first. This can happen when X substitutes one theory for another in his corpus.

No matter which kind of modification is made, I shall suppose that there is a "weakest" potential corpus UK (the "urcorpus") of sentences in L such that no rational agent should contract that corpus. UK is the deductively closed set of sentences in L such that every potential corpus in L is an expansion of UK (or is UK itself). I shall suppose that UK contains logical truths, set-theoretical truths, mathematical truths, and whatever else might be granted immunity from removal from the status of knowledge. (The items in UK are in this sense incorrigible.)

Replacement poses special problems for an account of the revision of knowledge. At t when X's corpus is $K_{X,t}$, why should he shift to a corpus K^* which is obtained by deleting items from $K_{X,t}$ and replacing them with other items inconsistent with the first? From X's point of view, at t, he is replacing a theory which he is certain is true by another which he is certain is false.

The puzzle can be avoided by regarding replacements for purposes of analysis as involving two steps: (a) contraction to a corpus relative to which no question is begged concerning the rival theories, and (b) subsequent expansion based on the information available in the contracted corpus, supplemented, perhaps, by the results of experiments conducted in the interim.

Those who insist on attempting to justify replacements without decomposing them into contractions followed by expansions confront the predicament that they cannot justify such shifts without begging questions. Such justification is no justification. The conclusion that beckons is that all replacements are forms of "conversion" to which men are subjected under revolutionary stress. This is the view which Thomas Kuhn has made so popular and which stands opposed to views that look to the formulation of objective criteria for the evaluation of proposed modifications of knowledge.

III

How does all this relate to bayesian views about the revision of credal states?

Consider X's corpus of knowledge $K_{X,t}$ at t. X's credal state $B_{X,t}$ at t is, according to strict bayesians, determined by $K_{X,t}$. Strict bayesians disagree among themselves concerning the appropriate way in which to formulate this determination. The following characterization captures the orthodox view in all its essentials.

Let K be any potential corpus (i.e., let it be UK or an expansion thereof). Let $C_{X,t}(K)$ be X's judgment at t as to what his credal state should be were he to adopt K as his corpus of knowledge. I shall suppose that X is committed to judgments of this sort for every feasible K in L. The resulting function from potential corpora of knowledge to potential credal states shall be called X's "confirmational commitment" at t.

According to strict bayesians, no matter what corpus K is (provided it is consistent), $C_{X,t}(K)$ is representable by a probability function where all sentences in K receive probability 1. In particular, $C_{X,t}(UK)$ is representable by a function $P(x;y)$—which I shall call a P-function, to contrast it with a Q-function representing $C_{X,t}(K)$ where K is an expansion of UK.

Strict bayesians adopt the following principle, which imposes restrictions upon confirmational commitments:

Confirmational Conditionalization: If K is obtained from UK by adding e (consistent with UK) to UK and forming the deductive closure, $P(x;y)$ represents $C_{x,t}(UK)$ and $Q(x;y)$ represents $C_{x,t}(K)$, $Q(h;f) = P(h;f\&e)$

In virtue of this principle, X's confirmational commitment is defined by specifying $C_{X,t}(UK) = C_{X,t}$ and employing confirmational conditionalization.[3] X's credal state at t, $B_{X,t}$, is then determined by $K_{X,t}$ and $C_{X,t}$ according to the following principle:

Total Knowledge: $C_{x,t}(K_{x,t}) = B_{x,t}$

Notice that the principle of confirmational conditionalization, even when taken together with the principle of total knowledge, does not prescribe how X should modify his credal state given a change in his corpus of knowledge.

To see this, suppose that at t_1 X's corpus is K_1 and that at t_2 his corpus K_2 is obtained from K_1 by adding e (consistent with K_1) and forming the deductive

[3] Confirmational commitments built on the principle of confirmational conditionalization are called "credibilities" by Carnap (*ibid.*, pp. 17–19). The analogy is not quite perfect. According to Carnap, a credibility function represents a permanent disposition of X to modify his credal states in the light of changes in his corpus of knowledge. When credibility is rational, it can be represented by a "confirmation function." Since I wish to allow for modifications of confirmational commitments as well as bodies of knowledge and credal states, I assign dates to confirmational commitments. Throughout I gloss over Carnap's distinction between credibility functions and confirmation functions (*ibid.*, pp. 24–27).

closure. From confirmational conditionalization and total knowledge, we can conclude that *if X does not alter his confirmational commitment in the interim from* t_1 *to* t_2, then, if Q_1 represents B_{X,t_1} and Q_2 represents B_{X,t_2}, $Q_2(h;f) = Q_1(h;f\&e)$. Should X renege at t_2 on the confirmational commitment he adopted at t_1, the change in knowledge just described need not and will not, in general, lead to a modification of credal state of the sort indicated.

Nonetheless, strict bayesians unanimously suppose that a rational agent will, save under unusual circumstances, modify his credal state in the fashion indicated. This mode of revising credal states is often called "conditionalization"; to distinguish it from confirmational conditionalization and other types of conditionalization, I shall call it "intertemporal credal conditionalization." I contend that the strict bayesian endorsement of intertemporal credal conditionalization presupposes commitment to the following principle:

Confirmational Tenacity: For every X, t, and t', $C_{x,t} = C_{x,t'}$

Thus, strict bayesians have an answer to the problem of revising credal states. X's confirmational commitment is to be held fixed over time. Given such a fixed commitment, the credal state he should adopt is determined for each possible modification of his corpus of knowledge which is a consistent expansion of UK. The problem of revising credal states reduces, therefore, to the problem of revising corpora of knowledge.

Is this answer to the problem of revision satisfactory? It would be, in my opinion, if the program for inductive logic envisaged by Carnap in his early writings on the subject could be realized. Inductive logic would then be strong enough to single out a standard P-function that all rational agents should adopt as their confirmational commitment. A fortiori, all such agents should hold that commitment fast at all times.

Few bayesians now think an inductive logic of the requisite power can be constructed. Their reasons (which, in my opinion, are sound) need not detain us. In response to this skepticism, most bayesians no longer require that all rational agents endorse a single standard confirmational commitment. They hold that rational X is perfectly free to pick any confirmational commitment consonant with the principles of inductive logic. Rational Y is quite free to pick a different commitment. However, bayesians tend to insist that, once X and Y have chosen their respective commitments, they should hold them fixed. To do this is to follow the probabilistic analogue of the method of tenacity so justly criticized by Peirce in "Fixation of Belief."

In the spirit of Peirce, it would have been far better to say that a rational X should not modify his confirmational commitment capriciously—i.e., without justification. To follow this approach, however, demands consideration of criteria for justified modifications of confirmational commitments. Bayesians not only fail to do this, but, as I shall now argue, they cannot do so without great difficulty. Given the bayesian answer to the problem of rational credence, no shift can be justified. If I am right, for bayesians, either tenacity should be obeyed, or, if not, justification is gratuitous. I think this implication of bayesian doctrine is to be deplored and should lead to scrutiny of other approaches.

IV

Modifying a confirmational commitment is not quite the same as modifying a corpus of knowledge. Yet, shifting from a confirmational commitment represented by a precise probability function to another confirmational commitment represented by a different precise probability function seems analogous to replacement in the following sense: The shift from confirmational commitment C_1 to confirmational commitment C_2 involves a shift to a confirmational commitment conflicting with C_1 in the sense that the P-function X uses to determine his credal state relative to his corpus when C_1 is adopted yields different precise subjective probability or credence assignments for hypotheses from those which X would make were he to adopt C_2 (and keep his corpus constant).

From X's vantage point at t when he endorses C_1, C_2 is illegitimate. He cannot justify shifting to C_2. At least, he cannot justify a direct shift. Can he do so indirectly by first performing a shift analogous to contraction from C_1 to C_3, which begs no questions concerning the merits of C_1 and C_2? Not from a strict bayesian point of view; for C_3 would, like C_1 and C_2, have to be representable by a precise P-function. The shift from C_1 to C_3 would be as problematic as the shift from C_1 to C_2.

Thus, from a bayesian point of view, no shift from one confirmational commitment to another can be justified. A rational man should conform to confirmational tenacity so that no justification is needed or else hold that some shifts are permitted without justification. Carnap sometimes seems to recognize shifts in confirmational commitments as a result of conceptual change.[4] Alternatively, one might allow shifts in confirmational commitment due to conversion under revolutionary stress. Except for the minimal requirement that the shift be to a commitment obeying requirements of inductive logic, no critical control is to be exercised. Bayesians are committed to being dogmatically tenacious or arbitrarily capricious.

The source of the difficulty should be apparent. Bayesians restrict the confirmational commitments a rational agent may adopt to those representable by numerically precise probability functions. This precludes shifting from a confirmational commitment C_1 to a confirmational commitment C_3 that begs no questions as to the merits of C_1 and another commitment C_2 that conflicts with C_1. My thesis is that not only are rational men allowed to make shifts to non-question-begging commitments but that on many occasions they ought to do so. That is to say, it is sometimes appropriate for a rational agent to adopt a confirmational commitment that is indeterminate in the sense that it cannot be represented by a numerically precise probability function. If we relax the stringent requirements imposed by bayesians on confirmational commitments and credal states so as to allow for such shifts, there is at least some hope that we can avoid endorsement of tenacity or capriciousness. Within the strict bayesian framework, we cannot expect to do so except by clinging desperately to Carnap's early program for constructing

[4] "A Basic System of Inductive Logic," in Carnap and Jeffrey, *op. cit*., pp. 51–52.

an inductive logic so strong as to single out a standard *P*-function to represent the uniquely rational confirmational commitment (for a given language).

I propose to explore one way of relaxing strict bayesian requirements. The basic idea is to represent a credal state (confirmational commitment) by a *set* of *Q*-functions (*P*-functions). When the set is single-membered, the credal state (confirmational commitment) will be indistinguishable in all relevant respects from a strict bayesian credal state (confirmational commitment).

On this view, if X starts at t with the precise (i.e., single-membered) confirmational commitment C_1, he can then shift to a confirmational commitment that has as members all the *P*-functions in C_1 as well as the *P*-functions in some other confirmational commitment C_2. (As the technical formulation will indicate, other *P*-functions will be members of C_3 as well.)

C_3 will be "weaker" than C_1 or C_2 in that it will allow more *P*-functions to be "permissible" than either of the other two confirmational commitments alone does. It will allow as permissible all *P*-functions recognized as such according to C_1 and according to C_2. In this sense, the shift to C_3 will beg no questions as to the permissibility of the *P*-functions in the other two confirmational commitments.

Of course, the notion of a permissible *P*-function (and the correlative notion of a permissible *Q*-function according to a credal state) require elucidation. I shall offer only an indirect clarification. The account of rational credence (and confirmational commitment) based on the new proposal will be supplemented by criteria for rational choice which indicate how permissibility determines the admissibility of options. By indicating the connections between permissibility and rational choice, permissibility will have been characterized indirectly.

V

To simplify the technical details, I shall restrict the discussion to characterizing credal states and confirmational commitments for sentences in a given language L which belong to a set M generated as follows: Let $h_1, h_2, \ldots, h_n$ be a finite set of sentences in L all consistent with the urcorpus UK for L and such that UK logically implies the truth of at least and at most one h_i. M is the set of sentences in L which are equivalent, given UK, to a disjunction of zero or more distinct h_is. (A disjunction of zero h_is is, as usual, a sentence inconsistent with UK.)

With this understanding, X's credal state at t will be a set $B_{X,t}$ of functions $Q(x;y)$ where the sentences substituted for 'x' are in M and the sentences substituted for 'y' are in M and are consistent with $K_{X,t}$. When the sentence substituted for 'y' is a member of $K_{X,t}$, I shall write $Q(x) = Q(x;y)$.

The set $B_{X,t}$ must satisfy the following three conditions:

1. *Nonemptiness:* $B_{X,t}$ is nonempty.
2. *Convexity:* $B_{X,t}$ is a convex set—i.e., every weighted average of *Q*-functions in $B_{X,t}$ is in $B_{X,t}$.
3. *Coherence:* Every *Q*-function in $B_{X,t}$ is a probability measure where $Q(h;e) = 1$ if and only if h is deductively implied by e and $K_{X,t}$.

Every Q-function in $B_{X,t}$ is "permissible" according to $B_{X,t}$.

As before, X's confirmational commitment $C_{X,t}(K)$ is a function from feasible corpora of knowledge to potential credal states that X at t considers to be the credal states he should adopt were he to adopt K as his corpus of knowledge. The value of the function for given K, therefore, is a nonempty, convex set of Q-functions relative to K. $C_{X,t}(UK) = C_{X,t}$ is, therefore, a nonempty convex set of P-functions. The principle of confirmational conditionalization introduced previously must now be modified to conform to the new characterization of confirmational commitments and credal states:

Confirmational Conditionalization: Let K be obtained from UK by adding e (consistent with UK) to UK and forming the deductive closure. $C_{x,t}(K)$ is the set of Q-functions such that $Q(h;f) = P(h;f\&e)$ for some permissible P-function in $C_{x,t} = C_{x,t}(UK)$.

$B_{X,t}$ can be determined, as before, as follows:

Total Knowledge: $B_{x,t} = C_{x,t}(K_{x,t})$

Thus, X's confirmational commitment is defined by specifying the value of $C_{X,t}(UK)$.

A strict bayesian confirmational commitment, of course, allows a single P-function to be uniquely permissible. However, confirmational commitments are possible which contain more than one P-function. In general, I shall say that one confirmational commitment is stronger than another if the set of its P-functions is a subset of the set of P-functions in the other commitment.

On this view, the weakest confirmational commitment possible is that which contains all the P-functions that meet the requirements of inductive logic. I shall continue to follow Carnap in understanding inductive logic to be a system of principles that impose constraints on probability functions eligible for membership in confirmational commitments.

In contrast, the strongest confirmational commitment would be the empty one—which is inconsistent with our first requirement of nonemptiness. A strongest "consistent" confirmational commitment is single-membered.

We can, by the way, extend the notion of a confirmational commitment so as to define it for an inconsistent corpus. We can require that $C_{X,t}(K)$ where K is inconsistent, be empty. This means that our previous requirement that a credal state be nonempty is to be restricted to cases where K is consistent. Thus, X might adopt a consistent confirmational commitment (i.e., one that is nonempty). Yet, if he should, unfortunately, endorse an inconsistent K, his credal state should be empty.

As noted previously, strict bayesians have differed among themselves as to what constitutes a complete system of principles of inductive logic. These differences persist on the view I am now proposing. They may be viewed, however, in a new light. The disagreements over inductive logic turn out to be disagreements over what constitutes the "weakest" possible confirmational commitment—which I shall call "*CIL(UK)*."

"Coherentists" like de Finetti and Savage claim that the principle of coherence constitutes a complete inductive logic. On their view, *CIL*(*UK*) is the set of all *P*-functions obeying the calculus of probabilities defined over *M*.

Some authors are prepared to add a further principle to the principle of coherence. This principle determines permissible *Q*-values for hypotheses about the outcome of a specific experiment on a chance device, given suitable knowledge about the experiment to be performed and the chances of possible outcomes of experiments of that type.

There is considerable controversy concerning the formulation of such a principle of "direct inference." In large measure, the controversy reflects disagreements over the interpretation of "chance" or "statistical probability," concerning the so-called "problem of the reference class" and random sampling. Indeed, the reason coherentists do not endorse a principle linking objective chance with credence is that they either deny the intelligibility of the notion of objective chance or argue in favor of dispensing with that notion.

Setting these controversies to one side, I shall call anyone who holds that a complete inductive logic consists of the coherence principle and an additional principle of direct inference from knowledge of chance to outcomes of random experiments an "objectivist."

There are many authors who are neither coherentists nor objectivists because they wish to supplement the principles of coherence and direct inference with additional principles. Some follow J. M. Keynes, Jeffreys, and Carnap in adding principles of symmetry of various kinds. Others, like I. Hacking,[5] introduce principles of irrelevance or other criteria which attempt to utilize knowledge about chances in a manner different from that employed in direct inference. Approaches of this sort stem by and large from the work of R. A. Fisher. I lack a good tag for this somewhat heterogeneous group of viewpoints. They all agree, however, in denying that objectivist inductive logic is a complete inductive logic.

Attempting to classify the views of historically given authors concerning inductive logic is fraught with risk. I shall not undertake a tedious and thankless task of textual analysis in the vain hope of convincing the reader that many eminent authors have been committed to an inductive logic whether they have said so or not. Yet much critical insight into controversies concerning probability, induction, and statistical inference can be obtained by reading the parties to the discussion as if they were committed to some form of inductive logic. If I am right, far from being a dead issue, inductive logic remains very much alive and debated (at least implicitly) not only by bayesians of the Keynes-Jeffreys-Carnap persuasion but by objectivists (to whose number I think J. Neyman, H. Reichenbach, and, with some qualifications, H. Kyburg belong) and the many authors, like Hacking, who are associated with the tradition of Fisher in various ways.

Assuming, for the sake of the argument, that the debate concerning what constitutes a complete set of principles of inductive logic is settled (I, for one, would

[5] *Logic and Statistical Inference* (New York: Cambridge, 1965), p. 135.

defend and will defend elsewhere adopting a variant of an objectivist inductive logic), there is yet another dimension to debates among students of probability, induction, and statistical inference.

Some authors seem to endorse the view that a rational agent should adopt the weakest confirmational commitment, *CIL*, consonant with inductive logic and hold it fast. They are, in effect, advocating confirmational tenacity. They do so, however, on the grounds that one should not venture to endorse a confirmational commitment stronger than the weakest allowed by inductive logic. (Their view is analogous to one that would require adopting the weakest corpus of knowledge *UK* and holding it fast.) I shall call advocates of such a view "necessitarians."

Again, classifying historically given authors is a risky business. However, Keynes, Jeffreys, and Carnap (in his early work) seem to be clear examples of necessitarians. What is more interesting is the implication that anyone is a necessitarian who insists that the only conditions under which a numerically precise probability can be assigned to a statement (other than a statement that is certainly true or false) are those derivable via direct inference from knowledge of chances. Such authors, on my view, are committed to saying that, when numerical probabilities are not assignable in this way, any numerical value is a permissible assignment provided that it is derived from Q-functions allowed by inductive logic.

To illustrate, suppose that X knows that a given coin has a .4 or a .6 chance of landing heads on a toss. Let h_1 be the first hypothesis that the chance is .4, and h_2 the second hypothesis. Let g be the hypothesis that the coin will land heads on the next toss. By direct inference, every permissible Q-function in X's credal state must be such that $Q(g;h_1) = .4$ and $Q(g;h_2) = .6$. By coherence, every Q-function in his credal state must be such that $Q(h_2) = 1 - Q(h_1)$, where $Q(h_1)$ is some real number between 0 and 1 and $Q(g) = Q(g;h_1)Q(h_1) + Q(g;h_2)Q(h_2) = .4Q(h_1) + .6(1 - Q(h_1))$.

According to the authors I have in mind, there is no unique numerical value that a rational X should adopt as uniquely permissible for $Q(h_1)$. As I am interpreting such authors as Kyburg, Neyman, Reichenbach, and Salmon, they mean to say that X's credal state should consist of all Q-functions meeting the conditions specified. The upshot is that the set of permissible Q-values for g should consist of all Q-values in the interval from .4 to .6. If I am reading them right, they endorse an objectivist logic and, at the same time, insist that X should adopt *CIL* as his confirmational commitment. They are "objectivist necessitarians."

The early Carnap, as noted previously, had hoped to identify an inductive logic that singled out a unique P-function as eligible for membership in confirmational commitments. Had his hope been realized, a rational agent would perforce have had to be a necessitarian. The weakest confirmational commitment would have been the strongest consistent one as well. Confirmational tenacity would have been necessitated by the principles of inductive logic.

But if Carnap's program is abandoned, necessitarianism is by no means the only response that one can make. Indeed, it seems to be of doubtful tenability, if for no other reason than that credal states formed on a necessitarian basis seem to be too

weak for use in practical decision making or statistical inference. (Many objectivist necessitarians seem to deny this; but the matter is much too complicated to discuss here.)

Personalists, like de Finetti and Savage, abandon necessitarianism but continue to endorse confirmational tenacity—at least during normal periods free from revolutionary stress. It is this position that I contended earlier leads to dogmatism or capriciousness with respect to confirmational commitment.

The view I favor is *revisionism*. This view agrees with the personalist position in allowing rational men to adopt confirmational commitments stronger than *CIL*. It insists, however, that such commitments are open to revision. It sees as a fundamental epistemological problem the task of providing an account of the conditions under which such revision is appropriate and criteria for evaluating proposed changes in confirmational commitment on those occasions when such shifts are needed.

I shall not offer an account of the revision of confirmational commitments. The point I wish to emphasize here is that, once one abandons the strict bayesian approach to credal rationality and allows credal states to contain more than one permissible Q-function in the manner I am suggesting, the revisionist position can be seriously entertained. The strict bayesian view precludes it and leaves us with the dubious alternatives of necessitarianism and personalism. By relaxing the strict bayesian requirements on credal rationality, we can at least ask a question about revision which could not be asked before.

VI

According to the approach I am proposing, X's credal state at t is characterized by a set of Q-functions defined over sentences in a set M. Such a representation describes X's credal state globally. Nothing has been said thus far as to how individual sentences in M are to be assigned grades of credence or how the degrees of credence assigned to two or more sentences are to be compared with one another. The following definitions seem to qualify for this purpose:

Def. 1: $Cr_{x,t}(h;e)$ is the set of real numbers r such that there is a Q-function in $B_{x,t}$ according to which $Q(h;e) = r$.

Def. 2: $c_{x,t}(h;e)$ is the set of real numbers r such that there is a P-function in $C_{x,t}$ according to which $P(h;e) = r$.

In virtue of the convexity requirement, both the credence function $Cr_{x,t}(h;e)$ and the confirmation function $c_{x,t}(h;e)$ will take sets of values that are subintervals of the unit line—i.e., the interval from 0 to 1. The lower and upper

bounds of such intervals have properties which have been investigated by I. J. Good,[6] C. A. B. Smith,[7] and A. P. Dempster.[8]

A partial ordering with respect to comparative credence or with respect to comparative confirmation can be defined as follows:

Def. 3: $(h;e) \overset{Cr_{x,t}}{\leq} (h';e')$ if and only if, for every Q-function in $B_{x,t}$, $Q(h;e) \leq Q(h';e')$.

Def. 4: $(h;e) \overset{C_{x,t}}{\leq} (h';e')$ if and only if, for every P-function in $C_{x,t}$, $P(h;e) \leq P(h';e')$.

The partial orderings induced by credal states and confirmational commitments conform to the requirements of B. O. Koopman's axioms for comparative probability.[9] Koopman pioneered in efforts to relax the stringent requirements imposed by bayesians on rational credence. Within the framework of his system, he was able not only to specify conditions of rational comparative probability judgment but to identify ways of generating interval-valued credence functions.

According to Koopman's approach, however, any two credal states (confirmational commitments) represented by the same partial ordering of the elements of M are indistinguishable. My proposal allows for important differences. Several distinct convex sets of probability distributions over the elements of M can induce the same partial ordering on the elements of M according to definitions 3 and 4.

Dempster, Good, Kyburg, Smith, and F. Schick, have all proposed modifying bayesian doctrine by allowing credal states and confirmational commitments to be represented by interval-valued probability functions.[10] Good, Smith, and Dempster have also explored the representation of credal states defined by interval-valued credence functions by means of sets of probability measures. Smith and Dempster explicitly consider convex sets of measures. Nonetheless, all these authors, including Dempster and Smith, seem to regard credal states (and confirmational commitments) represented by the same interval-valued function as indistinguishable. In contrast, my proposal recognizes credal states as different

[6] "Subjective Probability as the Measure of a Non-measurable Set," in P. Suppes, E. Nagel, and A. Tarski, *Logic*, *Methodology*, *and the Philosophy of Science* (Stanford: University Press, 1962), pp. 319–329.

[7] "Consistency in Statistical Inference and Decision" (with discussion), *Journal of the Royal Statistical Society*, series B, XXIII (1961): 1–25.

[8] "Upper and Lower Probabilities Induced by a mutivalued Mapping," *Annals of Mathematical Statistics*, XXXVIII (1967): 325–339.

[9] "The Bases of Probability," *Bulletin of the American Mathematical Society*, XLVI (1940): 763–774.

[10] Dempster, *op. cit.*; Good, *op. cit.*; Kyburg, *Probability and the Logic of Rational Belief* (Middletown, Conn.: Wesleyan Univ. Press, 1961): Smith, *op. cit.*; Schick, *Explication and Inductive Logic*, doctoral dissertation, Columbia University, 1958.

even though they generate the identical interval-valued function—provided they are different convex sets of Q-functions.[11]

Thus, the chief difference between my proposal and other efforts to come to grips with "indeterminate" probability judgments is that my proposal recognizes significant differences between credal states (confirmational commitments) where other proposals recognize none. Is this a virtue, or are the fine distinctions allowed by my proposal so much excess conceptual baggage?

I think that the distinctions between credal states recognized by the proposals introduced here are significant. Agents X and Y, who confront the same set of feasible options and evaluate the possible consequences in the same way may, nonetheless, be obliged as rational agents to choose different options if their credal states are different, even though their credal states define the same interval-valued credence function. That is to say, according to the decision theory that supplements the account of rational credence just introduced, differences in credal states recognized by my theory but not by Dempster's or Smith's, do warrant different choices in otherwise similar contexts of choice.

To explain this claim, we must turn to a consideration of rational choice. We would have to do so anyhow. One of the demands that can fairly be made of those who propose theories rival to bayesianism is that they furnish answers not only to the problems of rational credence and revision but to the questions about rational choice. Furthermore, the motivation for requiring credal states to be non-empty, convex sets of probability measures and the explanation of the notion of a permissible Q-function are best understood within the context of an account of rational choice. For all these reasons, therefore, it is time to discuss rational choice.

VII

Consider, once more, a situation where X faces a decision problem of the type described in section "I". No longer, however, will it be supposed that X's credal state for the "states of nature" $h_1, h_2, \ldots, h_n$ and for the possible consequences o_{i1} $o_{i2}, \ldots, o_{im}$ conditional on X choosing A_i are representable by a single Q-function. Instead, the credal state will be required only to be a nonempty convex set of Q-functions.[12]

[11] The difference between my approach and Smith's was drawn to my attention by Howard Stein. To all intents and purposes, both Dempster and Smith represent credal states by the largest convex sets that generate the interval-valued functions characterizing those credal states. Dempster (332/3) is actually more restrictive than Smith. Dempster, by the way, wrongly attributes to Smith the position I adopt. To my knowledge, Dempster is the first to consider this position in print—even if only to misattribute it to Smith.

[12] As in section "I", I am supposing that "states of nature" are "independent" of options in the sense that, for every permissible Q-function, $Q(h_j) = Q(o_{ij};A_i)$. I have done this to facilitate the exposition. No question of fundamental importance is, in my opinion, thereby seriously altered.

Although I have not focused attention here on the dubiety of requiring X's evaluations of the o_{ij}s to be representable by a utility function unique up to a linear transformation, I do believe that rational men can have indeterminate preferences and will, for the sake of generality, relax the bayesian requirement as follows: X's system of evaluations of the possible consequences of the feasible options is to be represented by a set G of "permissible" u-functions defined over the o_{ij}s which is (a) nonempty, (b) convex, and such that all linear transformations of u-functions in G are also in G. A bayesian G is, in effect, such that all u-functions in it are linear transformations of one another. It is this latter requirement that I am abandoning.

In those situations where X satisfies strict bayesian conditions so that his credal state contains only a single Q-function and G contains all and only those u-functions which are linear transformations of some specific u-function u_1, an admissible option A_i is, according to the principle of maximizing expected utility, an option that bears maximum expected utility $E(A_i) = \sum_{i=1}^{m} Q(h_i)\, u_1(o_{ij})$. Notice that, if any linear transformation of u_1 is substituted for u_1 in the computation of expected utility, the ranking of options with respect to expected utility remains unaltered. Hence we can say that, according to strict bayesians, an option is admissible if it bears maximum expected utility relative to the uniquely permissible Q-function and to any of the permissible u-functions in G (all of which are linear transformations of u_1).

There is an obvious generalization of this idea applicable to situations where $B_{X,t}$ contains more than one permissible Q-function and G contains u-functions that are not linear transformations of one another. I shall say that A_i is *E-admissible* if and only if there is at least one Q-function in $B_{X,t}$ and one u-function in G such that $E(A_i)$ defined relative to that Q-function and u-function is a maximum among all the feasible options. The generalization I propose is the following:

E-admissibility: All admissible options are E-admissible.

The principle of E-admissibility is by no means novel. I. J. Good, for example, endorsed it at one time. Indeed, Good went further than this. He endorsed the converse principle that all E-admissible options are admissible as well.[13]

I disagree with Good's view on this. When X's credal state and goals select more than one option as E-admissible, there may be and sometimes are other considerations than E-admissibility which X, as a rational agent, should employ in choosing between them.

There are occasions where X identifies two or more options as E-admissible and where, in addition, he has the opportunity to defer decision between them. If that opportunity is itself E-admissible, he should as a rational agent "keep his options open." Notice that in making this claim I am not saying that the option of deferring choice between the other E-admissible options is "better" than the

[13] "Rational Decisions," *Journal of the Royal Statistical Society*, Ser. B, XIV (1952): 114.

other *E*-admissible options relative to *X*'s credence and values and the assessments of expected utility based thereon. In general, *E*-admissible options will not be comparable with respect to expected utility (although sometimes they will be). The injunction to keep one's options open is a criterion of choice that is based not on appraisals of expected utility but on the "option-preserving" features of options. Deferring choice is better than the other *E*-admissible options in this respect, but not with respect to expected utility.

Thus, a *P-admissible* option is an option that is (a) *E*-admissible and (b) "best" with respect to *E*-admissible option preservation among all *E*-admissible options. I shall not attempt to provide an adequate explication of clause (b) here. In the subsequent discussion, I shall consider situations where there are no opportunities to defer choice. Nonetheless, it is important to notice that, given a suitably formulated surrogate for (b), the following principle holds:

P-admissibility: All admissible options are *P*-admissible.

My disagreement with Good goes still further than this; for I reject not only the converse of *E*-admissibility but that of *P*-admissibility as well.

To illustrate, consider a situation that satisfies strict bayesian requirements. *X* knows that a coin with a .5 chance of landing heads is to be tossed once. *g* is the hypothesis that the coin will land heads. Under the circumstances, we might say that *X*'s credal state is such that all permissible *Q*-functions assign *g* the value $Q(g) = .5$. Suppose that *X* is offered a gamble on *g* where *X* gains a dollar if *g* is true and loses one if *g* is false. (I shall assume that *X* has neither a taste for nor an aversion to gambling and that, for such small sums, money is linear with utility). He has two options: to accept the gamble and to reject it. If he rejects it, he neither gains nor loses.

Under the circumstances described, the principle of maximizing expected utility may be invoked. It indicates that both options are optimal and, hence, in my terms *E*-admissible. Since there are no opportunities for delaying choice, both options (on a suitably formulated version of *P*-admissibility) become *P*-admissible.

Bayesians—and Good would agree with this—tend to hold that rational *X* is free to choose either way. Not only are both options *E*-admissible. They are both admissible. Yet, in my opinion, rational *X* should refuse the gamble. The reason is not that refusal is better in the sense that it has higher expected utility than accepting the gamble. The options come out equal on this kind of appraisal. Refusing the gamble is "better," however, with respect to the security against loss it furnishes *X*. If *X* refuses the gamble, he loses nothing. If he accepts the gamble, he might lose something. This appeal to security does not carry weight, in my opinion, when accepting the gamble bears higher expected utility than refusing it. However, in that absurdly hypothetical situation where they bear precisely the same expected utility, the question of security does become critical.

These considerations can be brought to bear on the more general situation where two or more options are *E*-admissible (even though they are not equal with respect to expected utility) and where the principle of *P*-admissibility does not weed out any options.

An *S*-admissible option (i.e., option admissible with respect to security) is an option that is *P*-admissible and such that there is a permissible *u*-function in *G* relative to which the minimum *u*-value assigned a possible consequence o_{ij} of option A_i is a maximum among all *P*-admissible options.[14]

S-admissibility: All admissible options are *S*-admissible.

I cannot think of additional criteria for admissibility which seem adequate. (But then I have no precise conditions of adequacy.) I think, perhaps, we should keep an open mind on this matter. Nonetheless, for the present, I shall tentatively assume that the converse of *S*-admissibility holds. This assumption will not alter the main course of the subsequent argument.

Even without detailed exploration of the ramifications of this decision theory, some of its main features are immediately apparent. It conforms to the strict bayesian injunction to maximize expected utility in those situations where *X* has a precise credal state and *G* contains *u*-functions that are all linear transformations of one another. In this sense, bayesian decision theory is a special case of mine.

Similarly, the proposed decision theory identifies situations where the well-known maximin criterion is applied legitimately. Customarily maximin is used to select that option from among all the *feasible* options which maximizes the minimum gain. This recommendation is legitimate, according to my theory, provided (1) *G* contains all and only *u*-functions that are linear transformations of one another, and (2) all feasible options are *P*-admissible. But even if condition (1) is satisfied, it can be the case that the maximin solution from among all the feasible options is not itself *E*-admissible and so cannot be considered to be *S*-admissible.

Finally, my proposal is able to discriminate between and cover a wider variety of situations where neither maximizing expected utility nor maximining can be invoked with much plausibility. Moreover, it does so with the aid of a unified system of criteria of rational credence and rational choice. Thus, it does offer answers to just those questions which Bayesian theory purports to solve. Move rover, it escapes the bayesian commitment to the dubious doctrines of necessitarianism or personalism.

[14]The possible consequences of a "mixed act" constructed by choosing between "pure options" A_i and A_j with the aid of a chance device with known chance probability of selecting one or the other option is the set of possible consequences of either A_i or A_j. Consequently, the security level of such a mixed option for a given *u*-function is the lowest of the security levels belonging to A_i and A_j. Thus, my conception of security levels for mixed acts differs from that employed by von Neumann and Morgenstern and by Wald in formulating maximin (or minimax) principles. For this reason, starting with a set of *P*-admissible pure options, one cannot increase the security level by forming mixtures of them. In any case, mixtures of *E*-admissible options are not always themselves *E*-admissible. I shall leave mixed options out of account in the subsequent discussion. See D. Luce and H. Raiffa, *Games and Decisions* (New York: Wiley, 1958), pp. 68–71, 74–76, 278–280.

VIII

Some elementary properties of credal states as nonempty convex sets will be illustrated and explained by applying the decision theory just outlined to simple gambling situations. Suppose X knows that a coin is to be tossed and has either a .4 or .6 chance of landing heads. g is the hypothesis that the coin will land heads. I shall suppose that X has neither a taste for nor an aversion to gambling and that X's values are such that G is a set of u-functions that are linear transformations of the monetary payoffs of the gambles to be considered.

Case 1: X is offered a gamble on a take-it-or-leave-it basis where he receives $S - P$ dollars if g is true and loses P dollars if g is false. (Both S and P are positive.)

Case 2: X is offered a gamble on a take-it-or-leave-it basis where he loses P dollars if g is true and receives $S - P$ dollars if g is false. (S and P have the same values as in case 1.)

h_1 is the hypothesis that the chance of heads is .4, and h_2 is the hypothesis that the chance of heads is .6. By the reasoning of page 455, every permissible Q-function in X's credal state should be such that $Q(g) = .4Q(h_1) + .6[1 - Q(h_1)]$.

According to strict bayesians, X should, therefore, adopt a credal state that selects a single such Q-function as permissible. This can be done by selecting a single value for $Q(h_1)$. If that value is r, $Q(g) = .4r + .6(1 - r) = .6 - .2r$.

Hence, the bayesian will find that accepting the case 1 gamble is uniquely admissible if and only if $Q(g) > P/S$, and will find accepting the case 2 gamble uniquely admissible if and only if $Q(g) > P/S$. (Otherwise rejecting the gamble for the appropriate case is uniquely admissible, assuming that ties in expected utility are settled in favor of rejection.) Hence, if P/S is less than .5, a bayesian must preclude the possibility of accepting the gamble being inadmissible both in case 1 and in case 2.

Suppose, however, that $Cr_{X,t}(h_1)$ takes a nondegenerate interval as a value. For simplicity, let that interval be [0, 1]. The set of permissible Q-values for g must be all values of $.6 - .2r$ where r takes any value from 0 to 1. Hence, $Cr_{X,t}(g) = [.4, .6]$.

Under these conditions, my proposal holds that, when P/S falls in the interval from .4 to .6, both options are E-admissible (and P-admissible) in case 1. The same is true in case 2. But in both case 1 and case 2, rejecting the gamble is uniquely S-admissible. Hence, in both cases, X should reject the gamble. *This is true even when P/S is less than .5*. In this case, my proposal allows a rational agent a system of choices that a strict bayesian would forbid. In adopting this position, I am following the analysis advocated by C. A. B. Smith for handling pairwise choices between accepting and rejecting gambles. Smith's procedure, in brief, is to characterize X's degree of credence for g by a pair of numbers (the "lower pignic probability" and the "upper pignic probability" for g) as follows: The lower pignic probability $\underline{s}$ represents the least upper bound of betting quotients P/S for which X is prepared to accept gambles on g for positive S. The upper pignic probability $\overline{s}$ for g is $1 - t$, where t is the least upper bound of betting quotients P/S for which X is prepared to accept gambles on $\sim g$ for positive S. Smith requires that $\underline{s} \leq \overline{s}$, but does not insist

on equality as bayesians do. Given Smith's definitions of upper and lower pignic probabilities, it should be fairly clear that, in case 1 and case 2 where $Cr_{X,t}(g) = [.4, .6]$, Smith's analysis and mine coincide.[15]

Before leaving cases 1 and 2, it should be noted that, if X's credal state were empty, no option in case 1 would be admissible and no option in case 2 would be admissible either. If X is confronted with a case 1 predicament and an empty credal state, he would be constrained to act and yet as a rational agent enjoined not to act. The untenability of this result is to be blamed on adopting an empty credal state. Only when X's corpus is inconsistent, should a rational agent have an empty credal state. But, of course, if X finds his corpus inconsistent, he should contract to a consistent one.

Case 3: A_1 is accepting both the case 1 and the case 2 gamble jointly with a net payoff if g is true or false of $S - 2P$.

This is an example of decision making under certainty. Everyone agrees that if P is greater than $2S$ the gamble should be rejected; for it leads to certain loss. If P is less than $2S$ X should accept the gamble; for it leads to a certain gain. These results, by the way, are implied by the criteria proposed here as well as by the strict bayesian view.

Strict bayesians often defend requiring that Q-functions conform to the requirements of the calculus of probabilities by an appeal to the fact that, when credal states contain but a single Q-function, a necessary and sufficient condition for having credal states that do not license sure losses (Dutch books) is having a Q-function obeying the calculus of probabilities. The arguments also support the conclusion that, even when more than one Q-function is permissible according to a credal state, if all permissible Q-functions obey the coherence principle, no Dutch book can become E-admissible and, hence, admissible.

Case 4: B_1 is accepting the case 1 gamble, B_2 is accepting the case 2 gamble, and B_3 is rejecting both gambles.

[15] Smith, *op. cit*, pp. 3–5, 6–7. The agreement applies only to pairwise choices where one option is a gamble in which there are two possible payoffs and the other is refusing to gamble with 0 gain and 0 loss. In this kind of situation, it is clear that Smith endorses the principle of E-admissibility, but not its converse. However, in the later sections of his paper where Smith considers decision problems with three or more options or where the possible consequences of an option to be considered are greater than 2, Smith seems (but I am not clear about this) to endorse the converse of the principle of E-admissibility—counter to the analysis on the basis of which he defines lower and upper pignic probabilities. Thus, it seems to me that either Smith has contradicted himself or (as is more likely) he simply does not have a general theory of rational choice. The latter sections of the paper may then be read as interesting explorations of technical matters pertaining to the construction of such a theory, but not as actually advocating the converse of E-admissibility. At any rate, since it is the theory Smith propounds in the first part of his seminal essay which interests me, I shall interpret him in the subsequent discussion as having no general theory of rational choice beyond that governing the simple gambling situations just described.

Let the credal state be such that all values between 0 and 1 are permissible Q-values for h_1 and, hence, all values between .4 and .6 are permissible for g.

If P/S is greater than .6, B_3 is uniquely E-admissible and, hence, admissible. If P/S is less than .4, B_3 is E-inadmissible. The other two options are E-admissible and admissible.

If P/S is greater than or equal to .4 and less than .5, B_3 remains inadmissible and the other two admissible.

If P/S is greater than or equal to .5 and less than .6, all three options are E-admissible; but B_3 *is* uniquely S-admissible. Hence, B_3 should be chosen when P/S is greater than or equal to .5.

Three comments are worth making about these results.

(i) I am not sure what analysis Smith would propose of situations like case 4. At any rate, his theory does not seem to cover it (but see footnote 15).

(ii) When P/S is between .4 and .5, my theory recommends rejecting the gamble in case 1, rejecting the gamble in case 2, and yet recommends accepting one or the other of these gambles in case 4. This violates the so-called "principle of independence of irrelevant alternatives."[16]

(iii) If the convexity requirement for credal states were violated by removing as permissible values for g all values from $(S - P)/S$ to P/S, where P/S is greater than .5 and less than .6, but leaving all other values from .4 to .6, then—counter to the analysis given previously, B_3 would not be E-admissible in case 4. The peculiarity of that result is that B_1 is E-admissible because, for permissible Q-values from .6 down to P/S, it bears maximum expected utility, with B_3 a close second. B_2 is E-admissible because, for Q-values from .4 to $(S - P)/S$, B_2 is optimal, with B_3 again a close second. If the values between $(S - P)/S$ and P/S are also permissible, B_3 is E-admissible because it is optimal for those values. To eliminate such intermediate values and allow the surrounding values to retain their permissibility seems objectionable. Convexity guarantees against this.

Case 5: X is offered a gamble on a take-it-or-leave-it basis in which he wins 15 cents if f_1 is true, loses 30 cents if f_2 is true, and wins 40 cents if f_3 is true.

[16] See Luce and Raiffa, *op. cit.*, pp. 288/9. Because the analysis offered by Smith and me for cases 1 and 2 seems perfectly appropriate and the analysis for case 4 also appears impeccable, I conclude that there is something wrong with the principle of independence of irrelevant alternatives.

A hint as to the source of the trouble can be obtained by noting that if 'E-admissible' is substituted for 'optimal' in the various formulations of the principle cited by Luce and Raiffa, p. 289, the principle of independence of irrelevant alternatives stands. The principle fails because S-admissibility is used to supplement E-admissibility in weeding out options from the admissible set.

Mention should be made in passing that even when 'E-admissible' is substituted for 'optimal' in Axiom 9 of Luce and Raiffa, p. 292, the axiom is falsified. Thus, when $.5 \leq P/S \leq .6$ in case 4, all three options are E-admissible, yet some mixtures of B_1 and B_2 will not be.

Suppose X's corpus of knowledge contains the following information:

Situation a: X knows that the ratios of red, white, and blue balls in the urn are either (i) 1/8, 3/8, 4/8 respectively; (ii) 1/8, 4/8, 3/8; (iii) 2/8, 4/8, 2/8; or (iv) 4/8, 3/8, 1/8.

X's credal state for the f_is is determined by his credal state for the four hypotheses about the contents of the urn according to a more complex variant of the arguments used to obtain credence values for g in the first four cases. If we allow all Q-functions compatible with inductive logic of an objectivist kind to be permissible, X's credal state for the f_is is the convex set of all weighted averages of the four triples of ratios. $Cr_{X,t}(f_1) = (1/8,\ 4/8)$, $Cr_{X,t}(f_2) = (3/8,\ 4/8)$, and $Cr_{X,t}(f_3) = (1/8,\ 4/8)$. Both accepting and rejecting the gamble are E-admissible. Rejecting the gamble, however, is uniquely S-admissible. X should reject the gamble.

Situation b: X knows that the ratios of red, white, and blue balls is correctly described by (i), (ii), or (iv), but not by (iii). Calculation reveals that the interval-valued credence function is the same as in situation a. Yet it can be shown that accepting the gamble is uniquely E-admissible and, hence, admissible. X should accept the gamble.

Now we can imagine situations that are related as a and b are to one another except that the credal states do not reflect differences in statistical knowledge. Then, from the point of view of Dempster and Smith, the credal states would be indistinguishable. Because the set of permissible Q-distributions over the f_is would remain different for situations a and b, my view would recognize differences and recommend different choices. If the answer to the problem of rational choice proposed here is acceptable, the capacity of the account of credal rationality to make fine distinctions is a virtue rather than a gratuitous piece of pedantry.

The point has its ramifications for an account of the improvement of confirmational commitments; the variety of discriminations that can be made between confirmational commitments generates a variety of potential shifts in confirmational commitments subject to critical review. For intervalists, a shift from situation a to b is no shift at all. On the view proposed here, it is significant.

The examples used in this section may be used to illustrate one final point. The objective or statistical or chance probability distributions figuring in chance statements can be viewed as assumptions or hypotheses. Probabilities in this sense can be unknown. We can talk of a set of simple or precise chance distributions among which X suspends judgment. Such *possible* probability distributions represent hypotheses which are possibly true and which are themselves objects of appraisal with respect to credal probability. *Permissible* probability distributions which, in our examples, are defined over such *possible* probability distributions (like the hypotheses h_1 and h_2 of cases 1, 2, 3, and 4) are not themselves possibly true hypotheses. No probability distributions of a still higher type can be denned over them.[17]

[17]I mention this because I. J. Good, whose seminal ideas have been an important influence on the proposals made in this essay, confuses permissible with possible probabilities. As a consequence, he introduces a hierarchy of types of probability (Good, *op. cit*., p. 327). For criticism of such

I have scratched the surface of some of the questions raised by the proposals made in this essay. Much more needs to be done. I do believe, however, that these proposals offer fertile soil for cultivation not only by statisticians and decision theorists but by philosophers interested in what, in my opinion, ought to be the main problem for epistemology—to wit, the improvement (and, hence, revision) of human knowledge and belief.

Acknowledgements Work on this essay was partially supported by N.S.F. grant GS 28992. Research was carried out while I was a Visiting Scholar at Leckhampton, Corpus Christi, Cambridge. I wish to thank the Fellows of Corpus Christi College and the Departments of Philosophy and History and Philosophy of Science, Cambridge University, for their kind hospitality. I am indebted to Howard Stein for his help in formulating and establishing some of the results reported here. Sidney Morgenbesser, Ernest Nagel, Teddy Seidenfeld, and Frederic Schick as well as Stein have made helpful suggestions.

views, see Savage, *The Foundations of Statistics* (New York: Wiley, 1954), p. 58. In fairness to Good, it should be mentioned that his possible credal probabilities are interpreted not as possibly true statistical hypotheses but as hypotheses entertained by X about his own unknown strictly bayesian credal state. Good is concerned with situations where strict bayesian agents having precise probability judgments cannot identify their credal states before decision and must make choices on the basis of partial information about themselves. [In *Decision and Value* (New York: Wiley, 1964), P. G. Fishburn devotes himself to the same question.] My proposals do not deal with this problem. I reject Good's and Fishburn's view that every rational agent is at bottom a strict bayesian limited only by his lack of self-knowledge, computational facility, and memory. To the contrary, I claim that, even without such limitations, rational agents should not have precise bayesian credal states. The difference in problem under consideration and presuppositions about rational agents has substantial technical ramifications which cannot be developed here.

Chapter 8
Why I am not a Bayesian

Clark Glymour

The aim of confirmation theory is to provide a true account of the principles that guide scientific argument in so far as that argument is not, and does not purport to be, of a deductive kind. A confirmation theory should serve as a critical and explanatory instrument quite as much as do theories of deductive inference. Any successful confirmation theory should, for example, reveal the structure and fallacies, if any, in Newton's argument for universal gravitation, in nineteenth-century arguments for and against the atomic theory, in Freud's arguments for psychoanalytic generalizations. Where scientific judgements are widely shared, and sociological factors cannot explain their ubiquity, and analysis through the lens provided by confirmation theory reveals no good explicit arguments for the judgements, confirmation theory ought at least sometimes to suggest some good arguments that may have been lurking misperceived. Theories of deductive inference do that much for scientific reasoning in so far as that reasoning is supposed to be demonstrative. We can apply quantification theory to assess the validity of scientific arguments, and although we must almost always treat such arguments as enthymematic, the premisses we interpolate are not arbitrary; in many cases, as when the same subject-matter is under discussion, there is a common set of suppressed premisses. Again, there may be differences about the correct logical form of scientific claims; differences of this kind result in (or from) different formalizations, for example, of classical mechanics. But such differences often make no difference for the assessment of validity in actual arguments. Confirmation theory should do as well in its own domain. If it fails, then it may still be of interest for many purposes, but not for the purpose of understanding scientific reasoning.

Who cares whether a pig-farmer is a Bayesian?—R. C. Jeffrey.

C. Glymour (✉)
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: cg09@andrew.cmu.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_8

The aim of confirmation theory ought not to be simply to provide precise replacements for informal methodological notions, that is, explications of them. It ought to do more; in particular, confirmation theory ought to *explain* both methodological truisms and particular judgements that have occurred within the history of science. By 'explain' I mean at least that confirmation theory ought to provide a rationale for methodological truisms, and ought to reveal some systematic connections among them and, further, ought, without arbitrary or question-begging assumptions, to reveal particular historical judgements as in conformity with its principles.

Almost everyone interested in confirmation theory today believes that confirmation relations ought to be analysed in terms of *probability* relations. Confirmation theory is the theory of probability plus introductions and appendices. Moreover, almost everyone believes that confirmation proceeds through the formation of conditional probabilities of hypotheses on evidence. The basic tasks facing confirmation theory are thus just those of explicating and showing how to determine the probabilities that confirmation involves, developing explications of such meta-scientific notions as 'confirmation', 'explanatory power', 'simplicity', and so on in terms of functions of probabilities and conditional probabilities, and showing that the canons and patterns of scientific inference result. It was not always so. Probabilistic accounts of confirmation really became dominant only after the publication of Carnap's *Logical Foundations of Probability* (1950), although of course many probabilistic accounts had preceded Carnap's. An eminent contemporary philosopher (Putnam 1967) has compared Carnap's achievement in inductive logic with Frege's in deductive logic: just as before Frege there was only a small and theoretically uninteresting collection of principles of deductive inference, but after him the foundation of a systematic and profound theory of demonstrative reasoning, so with Carnap and inductive reasoning. After Carnap's *Logical Foundations*, debates over confirmation theory seem to have focused chiefly on the interpretation of probability and on the appropriate probabilistic explications of various meta-scientific notions. The meta-scientific notions remain controversial, as does the interpretation of probability, although, increasingly, logical interpretations of probability are giving way to the doctrine that probability is degree of belief.[1] In very recent years a few philosophers have attempted to apply probabilistic analyses to derive and to explain particular methodological practices and precepts, and even to elucidate some historical cases.

I believe these efforts, ingenious and admirable as many of them are, are none the less misguided. For one thing, probabilistic analyses remain at too great a distance from the history of scientific practice to be really informative about that practice, and in part they do so exactly because they are probabilistic. Although considerations of probability have played an important part in the history of science, until very recently, explicit probabilistic arguments for the confirmation of various

[1] A third view, that probabilities are to be understood exclusively as frequencies, has been most ably defended by Wesley Salmon (1969).

theories, or probabilistic analyses of data, have been great rarities in the history of science. In the physical sciences at any rate, probabilistic arguments have rarely occurred. Copernicus, Newton, Kepler, none of them give probabilistic arguments for their theories; nor does Maxwell or Kelvin or Lavoisier or Dalton or Einstein or Schrödinger or . . . There are exceptions. Jon Dorling has discussed a seventeenth-century Ptolemaic astronomer who apparently made an extended comparison of Ptolemaic and Copernican theories in probabilistic terms; Laplace, of course, gave Bayesian arguments for astronomical theories. And there are people—Maxwell, for example—who scarcely give a probabilistic argument when making a case for or against scientific hypotheses but who discuss *methodology* in probabilistic terms. This is not to deny that there are many areas of contemporary physical science where probability figures large in confirmation; regression analysis is not uncommon in discussions of the origins of cosmic rays, correlation and analysis of variance in experimental searches for gravitational waves, and so on. It *is* to say that, explicitly, probability is a distinctly minor note in the history of scientific argument.

The rarity of probability considerations in the history of science is more an embarrassment for some accounts of probability than for others. Logical theories, whether Carnap's or those developed by Hintikka and his students, seem to lie at a great distance from the history of science. Still, some of the people working in this tradition have made interesting steps towards accounting for methodological truisms. My own inclination is to believe that the interest such investigations have stems more from the insights they obtain into syntactic versions of structural connections among evidence and hypotheses than to the probability measures they mesh with these insights. Frequency interpretations suppose that for each hypothesis to be assessed there is an appropriate reference class of hypotheses to which to assign it, and the prior probability of the hypothesis is the frequency of true hypotheses in this reference class. The same is true for statements of evidence, whether they be singular or general. The matter of how such reference classes are to be determined, and determined so that the frequencies involved do not come out to be zero, is a question that has only been touched upon by frequentist writers. More to the point, for many of the suggested features that might determine reference classes, we have no statistics, and cannot plausibly imagine those who figure in the history of our sciences to have had them. So conceived, the history of scientific argument must turn out to be largely a history of fanciful guesses. Further, some of the properties that seem natural candidates for determining reference classes for hypotheses—simplicity, for example—seem likely to give perverse results. We prefer hypotheses that posit simple relations among observed quantities, and so on a frequentist view should give them high prior probabilities. Yet simple hypotheses, although often very useful approximations, have most often turned out to be literally false.

At present, perhaps the most philosophically influential view of probability understands it to be degree of belief. The subjectivist Bayesian (hereafter, for brevity, simply Bayesian) view of probability has a growing number of advocates who understand it to provide a general framework for understanding scientific reasoning. They are singularly unembarrassed by the rarity of explicit probabilistic

arguments in the history of science, for scientific reasoning need not be explicitly probabilistic in order to be probabilistic in the Bayesian sense. Indeed, a number of Bayesians have discussed historical cases within their framework. Because of its influence and its apparent applicability, in what follows it is to the subjective Bayesian account that I shall give my full attention.

My thesis is several-fold. First, there are a number of attempts to demonstrate a priori the rationality of the restrictions on belief and inference that Bayesians advocate. These arguments are altogether admirable, but ought, I shall maintain, to be unconvincing. My thesis in this instance is not a new one, and I think many Bayesians do regard these a priori arguments as insufficient. Second, there are a variety of methodological notions that an account of confirmation ought to explicate and methodological truisms involving these notions that a confirmation theory ought to explain: for example, variety of evidence and why we desire it, *ad hoc* hypotheses and why we eschew them, what separates a hypothesis integral to a theory from one 'tacked on' to the theory, simplicity and why it is so often admired, why 'de-Occamized' theories are so often disdained, what determines when a piece of evidence is relevant to a hypothesis, and what, if anything, makes the confirmation of one bit of theory by one bit of evidence stronger than the confirmation of another bit of theory (or possibly the same bit) by another (or possibly the same) bit of evidence. Although there are plausible Bayesian explications of some of these notions, there are not plausible Bayesian explications of others. Bayesian accounts of methodological truisms and of particular historical cases are of one of two kinds: either they depend on general principles restricting prior probabilities, or they don't. My claim is that many of the principles proposed by the first kind of Bayesian are either implausible or incoherent, and that, for want of such principles, the explanations the second kind of Bayesians provide for particular historical cases and for truisms of method are chimeras. Finally, I claim that there are elementary but perfectly common features of the relation of theory and evidence that the Bayesian scheme cannot capture at all without serious—and perhaps not very plausible—revision.

It is not that I think the Bayesian scheme or related probabilistic accounts capture nothing. On the contrary, they are clearly pertinent where the reasoning involved is explicitly statistical. Further, the accounts developed by Carnap, his predecessors, and his successors are impressive systematizations and generalizations, in a probabilistic framework, of certain principles of ordinary reasoning. But so far as understanding scientific reasoning goes, I think it is very wrong to consider our situation to be analogous to that of post-Fregean logicians, our subject-matter transformed from a hotchpotch of principles by a powerful theory whose outlines are clear. We flatter ourselves that we possess even the hotchpotch. My opinions are outlandish, I know; few of the arguments I shall present in their favour are new, and perhaps none of them is decisive. Even so, they seem sufficient to warrant taking seriously entirely different approaches to the analysis of scientific reasoning.

The theories I shall consider share the following framework, more or less. There is a class of sentences that express all hypotheses and all actual or possible evidence

of interest; the class is closed under Boolean operations. For each ideally rational agent, there is a function defined on all sentences such that, under the relation of logical equivalence, the function is a probability measure on the collection of equivalence classes. The probability of any proposition represents the agent's degree of belief in that proposition. As new evidence accumulates, the probability of a proposition changes according to Bayes's rule: the posterior probability of a hypothesis on the new evidence is equal to the prior conditional probability of the hypothesis on the evidence. This is a scheme shared by diverse accounts of confirmation. I call such theories 'Bayesian', or sometimes 'personalist'.

We certainly have *grades* of belief. Some claims I more or less believe, some I find plausible and tend to believe, others I am agnostic about, some I find implausible and far-fetched, still others I regard as positively absurd. I think everyone admits some such gradations, although descriptions of them might be finer or cruder. The personalist school of probability theorists claim that we also have *degrees* of belief, degrees that can have any value between 0 and 1 and that ought, if we are rational, to be representable by a probability function. Presumably, the degrees of belief are to co-vary with everyday gradations of belief, so that one regards a proposition as preposterous and absurd just if his degree of belief in it is somewhere near zero, and he is agnostic just if his degree of belief is somewhere near a half, and so on. According to personalists, then, an ideally rational agent always has his degrees of belief distributed so as to satisfy the axioms of probability, and when he comes to accept a new belief, he also forms new *degrees* of belief by conditionalizing on the newly accepted belief. There are any number of refinements, of course; but that is the basic view.

Why should we think that we really do have *degrees* of belief? Personalists have an ingenious answer: people have them because we can measure the degrees of belief that people have. Assume that no one (rational) will accept a wager on which he expects a loss, but anyone (rational) will accept any wager on which he expects a gain. Then we can measure a person's degree of belief in proposition P by finding, for fixed amount v, the highest amount u such that the person will pay u in order to receive $u + v$ if P is true, but receive nothing if P is not true. If u is the greatest amount the agent is willing to pay for the wager, his expected gain on paying u must be zero. The agent's gain if P is the case is v; his gain if P is not the case is $-u$. Thus

$$v \cdot \text{prob}(P) + (-u) \cdot \text{prob}\,(\sim P) = 0.$$

Since prob $(\sim P) = 1 - \text{prob}(P)$, we have

$$\text{prob}(P) = u/\,(u + v)\,.$$

The reasoning is clear: any sensible person will act so as to maximize his expected gain; thus, presented with a decision whether or not to purchase a bet, he will make

the purchase just if his expected gain is greater than zero. So the betting odds he will accept determine his degree of belief.[2]

I think that this device really does provide evidence that we have, or can produce, degrees of belief, in at least some propositions, but at the same time it is evident that betting odds are not an unobjectionable device for the measurement of degrees of belief. Betting odds could fail to measure degrees of belief for a variety of reasons: the subject may not believe that the bet will be paid off if he wins, or he may doubt that it is clear what constitutes winning, even though it is clear what constitutes losing. Things he values other than monetary gain (or whatever) may enter into his determination of the expected utility of purchasing the bet: for example, he may place either a positive or a negative value on risk itself. And the very fact that he is offered a wager on P may somehow change his degree of belief in P.

Let us suppose, then, that we do have degrees of belief in at least some propositions, and that in some cases they can be at least approximately measured on an interval from 0 to 1. There are two questions: why should we think that, for rationality, one's degrees of belief must satisfy the axioms of probability, and why should we think that, again for rationality, changes in degrees of belief ought to proceed by conditionalization? One question at a time. In using betting quotients to measure degrees of belief, it was assumed that the subject would act so as to maximize *expected* gain. The betting quotient determined the degree of belief by determining the coefficient by which the gain is multiplied in case that P is true in the expression for the expected gain. So the betting quotient determines a degree of belief, as it were, in the *role* of a probability. But why should the things, degrees of belief, that play this role be probabilities? Supposing that we do choose those actions that maximize the sum of the product of our degrees of belief in each possible outcome of the action and the gain (or loss) to us of that outcome. Why must the degrees of belief that enter into this sum be probabilities? Again, there is an ingenious argument: if one acts so as to maximize his expected gain using a degree-of-belief function that is not a probability function, and if for every proposition there were a possible wager (which, if it is offered, one believes will be paid off if it is accepted and won), then there is a circumstance, a combination of wagers, that one would enter into if they were offered, and in which one would suffer a net loss whatever the outcome. That is what the Dutch-book argument shows; what it counsels is prudence.

Some of the reasons why it is not clear that betting quotients are accurate measures of degrees of belief are also reasons why the Dutch-book argument is not conclusive: there are many cases of propositions in which we may have degrees of belief, but on which, we may be sure, no acceptable wager will be offered us;

[2]More detailed accounts of means for determining degrees of belief may be found in Jeffrey (1965). It is a curious fact that the procedures that Bayesians use for determining subjective degrees of belief empirically are an instance of the general strategy described in Glymour 1981, ch. 5. Indeed, the strategy typically used to determine whether or not actual people behave as rational Bayesians involves the bootstrap strategy described in that chapter.

again, we may have values other than the value we place on the stakes, and these other values may enter into our determination whether or not to gamble; and we may not have adopted the policy of acting so as to maximize our expected gain or our expected utility: that is, we may save ourselves from having book made against us by refusing to make certain wagers, or combinations of wagers, even though we judge the odds to be in our favour.

The Dutch-book argument does not succeed in showing that in order to avoid absurd commitments, or even the possibility of such commitments, one must have degrees of belief that are probabilities. But it does provide a kind of justification for the personalist viewpoint, for it shows that if one's degrees of belief are probabilities, then a certain kind of absurdity is avoided. There are other ways of avoiding that kind of absurdity, but at least the personalist way is one such.[3]

One of the common objections to Bayesian theory is that it fails to provide any connection between what is inferred and what is the case. The Bayesian reply is that the method guarantees that, in the long run, everyone will agree on the truth. Suppose that B_i are a set of mutually exclusive, jointly exhaustive hypotheses, each with probability $B(i)$. Let $\overline{x}_r$ be a sequence of random variables with a finite set of values and conditional distribution given by $P\left(\overline{x}_r = x_r \middle| B_i\right) = \varepsilon\left(x_r \middle| B_i\right)$; then we can think of the values x_r as the outcomes of experiments, each hypothesis determining a likelihood for each outcome. Suppose that no two hypotheses have the same likelihood distribution; that is, for $i \neq j$ it is not the case that for all values x_r of $\overline{x}_r$, $\varepsilon(x_r|B_i) = \varepsilon(x_r|B_j)$, where the ε's are defined as above. Let $\overline{x}$ denote the first n of these variables, where x is a value of $\overline{x}$. Now imagine an observation of these n random variables. In Savage's words:

Before the observation, the probability that the probability given x of whichever element of the partition actually obtains will be greater than α is

$$\sum_i B(i) P\left(P\left(B_i \middle| x\right) > \alpha \middle| B_i\right),$$

where summation is confined to those i's for which $B(i) \neq 0$. (1972: 49)

In the limit as n approaches infinity, the probability that the probability given x of whichever element of the partition actually obtains is greater than α is 1. That is the theorem. What is its significance? According to Savage, 'With the observation of an abundance of relevant data, the person is almost certain to become highly convinced of the truth, and it has also been shown that he himself knows this to be the case' (p. 50). That is a little misleading. The result involves second-order probabilities, but these too, according to personalists, are degrees of belief. So what has been shown seems to be this: in the limit as n approaches infinity, an ideally rational Bayesian has degree of belief 1 that an ideally rational Bayesian (with degrees of belief as in the theorem) has degree of belief, given x, greater than α

[3] For further criticisms of the Dutch-book argument see Kyburg 1978.

in whichever element of the partition actually obtains. The theorem does not tell us that in the limit any rational Bayesian will assign probability 1 to the true hypothesis and probability 0 to the rest; it only tells us that rational Bayesians are certain that he will. It may reassure those who are already Bayesians, but it is hardly grounds for conversion. Even the reassurance is slim. Mary Hesse points out (1974: 117–19), entirely correctly I believe, that the assumptions of the theorem do not seem to apply even approximately in actual scientific contexts. Finally, some of the assumptions of stable estimation theorems can be dispensed with if one assumes instead that all of the initial distributions considered must agree regarding which evidence is relevant to which hypotheses. But there is no evident a priori reason why there should be such agreement.

I think relatively few Bayesians are actually persuaded of the correctness of Bayesian doctrine by Dutch-book arguments, stable estimation theorems, or other a priori arguments. Their frailty is too palpable. I think that the appeal of Bayesian doctrine derives from two other features. First, with only very weak or very natural assumptions about prior probabilities, or none at all, the Bayesian scheme generates principles that seem to accord well with common sense. Thus, with minor restrictions, one obtains the principle that hypotheses are confirmed by positive instances of them; and, again, one obtains the result that if an event that actually occurs is, on some hypothesis, very unlikely to occur, then that occurrence renders the hypothesis less likely than it would otherwise have been. These principles, and others, can claim something like the authority of common sense, and Bayesian doctrine provides a systematic explication of them. Second, the restrictions placed a priori on rational degrees of belief are so mild, and the device of probability theory at once so precise and so flexible, that Bayesian philosophers of science may reasonably hope to explain the subtleties and vagaries of scientific reasoning and inference by applying their scheme together with plausible assumptions about the distribution of degrees of belief. This seems, for instance, to be Professor Hesse's line of argument. After admitting the insufficiency of the standard arguments for Bayesianism, she sets out to show that the view can account for a host of alleged features of scientific reasoning and inference. My own view is different: particular *inferences* can almost always be brought into accord with the Bayesian scheme by assigning degrees of belief more or less *ad hoc*, but we learn nothing from this agreement. What we want is an explanation of scientific argument; what the Bayesians give us is a theory of learning—indeed, a theory of personal learning. But arguments are more or less impersonal; I make an argument to persuade anyone informed of the premisses, and in doing so I am not reporting any bit of autobiography. To ascribe to me degrees of belief that make my slide from my premisses to my conclusion a plausible one fails to explain anything, not only because the ascription may be arbitrary, but also because, even if it is a correct assignment of my degrees of belief, it does not explain why what I am doing is *arguing*—why, that is, what I say should have the least influence on others, or why I might hope that it should. Now, Bayesians might bridge the gap between personal inference and argument in either of two ways. In the first place, one might give arguments in order to change others' beliefs because of the respect they have for his

opinion. This is not very plausible; if that were the point of giving arguments, one would not bother with them, but would simply state one's opinion. Alternatively, and more hopefully, Bayesians may suggest that we give arguments exactly because there are general principles restricting belief, principles that are widely subscribed to, and in giving arguments we are attempting to show that, supposing our audience has certain beliefs, they must in view of these principles have other beliefs, those we are trying to establish. There is nothing controversial about this suggestion, and I endorse it. What is controversial is that the general principles required for argument can best be understood as conditions restricting prior probabilities in a Bayesian framework. Sometimes they can, perhaps; but I think that when arguments turn on relating evidence to theory, it is very difficult to explicate them in a plausible way within the Bayesian framework. At any rate, it is worth seeing in more detail what the difficulties may be.

There is very little Bayesian literature about the hotchpotch of claims and notions that are usually canonized as scientific method; very little seems to have been written, from a Bayesian point of view, about what makes a hypothesis *ad hoc*, about what makes one body of evidence more various than another body of evidence, and why we should prefer a variety of evidence, about why, in some circumstances, we should prefer simpler theories, and what it is that we are preferring when we do. And so on. There is little to nothing of this in Carnap, and more recent, and more personalist, statements of the Bayesian position are almost as disappointing. In a lengthy discussion of what he calls 'tempered personalism', Abner Shimony (1970) discusses only how his version of Bayesianism generalizes and qualifies hypothetico-deductive arguments. (Shimony does discuss simplicity, but only to argue that it is overvalued.) Mary Hesse devotes the later chapters of her book to an attempt to show that certain features of scientific method do result when the Bayesian scheme is supplemented with a postulate that restricts assignments of prior probabilities. Unfortunately, as we shall see, her restrictive principle is incoherent.[4]

One aspect of the demand for a variety of evidence arises when there is some definite set of alternative hypotheses between which we are trying to decide. In such cases we naturally prefer the body of evidence that will be most helpful in eliminating false competitors. This aspect of variety is an easy and natural one for Bayesians to take account of, and within an account such as Shimony's it is taken care of so directly as hardly to require comment. But there is more to variety. In some situations we have some reason to suspect that if a theory is false, its falsity will show up when evidence of certain kinds is obtained and compared. For example, given the tradition of Aristotelian distinctions, there was some reason to demand both terrestrial and celestial evidence for seventeenth-century theories of motion that subjected all matter to the same dynamical laws. Once again, I see no special reason why this kind of demand for a variety of evidence cannot be fitted into the Bayesian scheme. But there is still more. A complex theory may contain

[4]Moreover, I believe that much of her discussion of methodological principles has only the loosest relation to Bayesian principles.

a great many logically independent hypotheses, and particular bodies of evidence may provide grounds for some of those hypotheses but not for others. Surely part of the demand for a variety of evidence, and an important part, derives from a desire to see to it that the various independent parts of our theories are tested. Taking account of this aspect of the demand for a variety of evidence is just taking account of the relevance of evidence to pieces of theory. How Bayesians may do this we shall consider later.

Simplicity is another feature of scientific method for which some Bayesians have attempted to account. There is one aspect of the scientific preference for the simple that seems beyond Bayesian capacities, and that is the disdain for 'de-Occamized' hypotheses, for theories that postulate the operation of a number of properties, determinable only in combination, when a single property would do. Such theories can be generated by taking any ordinary theory and replacing some single quantity, wherever it occurs in the statement of the theory, by an algebraic combination of new quantities. If the original quantity was not one that occurs in the statement of some body of evidence for the theory, then the new, de-Occamized theory will have the same entailment relations with that body of evidence as did the original theory. If the old theory entailed the evidence, so will the new, de-Occamized one. Now, it follows from Bayesian principles that if two theories both entail *e*, then (provided the prior probability of each hypothesis is neither 1 nor 0), if *e* confirms one of them, it confirms the other. How then is the fact (for so I take it to be) that pieces of evidence just don't seem to *count* for de-Occamized theories to be explained? Not by supposing that de-Occamized theories have lower prior probabilities than un-de-Occamized theories, for being 'de-Occamized' is a feature that a theory has only with respect to a certain body of evidence, and it is not hard to imagine artificially restricted bodies of evidence with respect to which perfectly good theories might count as de-Occamized. Having extra wheels is a feature a theory has only in relation to a body of evidence; the only Bayesian relation that appears available and relevant to scientific preference is the likelihood of the evidence on the theory, and unfortunately the likelihood is the same for a theory and for its de-Occamized counterparts whenever the theory entails the evidence.

It is common practice in fitting curves to experimental data, in the absence of an established theory relating the quantities measured, to choose the 'simplest' curve that will fit the data. Thus linear relations are preferred to polynomial relations of higher degree, and exponential functions of measured quantities are preferred to exponential functions of algebraic combinations of measured quantities, and so on. The problem is to account for this preference. Harold Jeffreys, a Bayesian of sorts, offered an explanation (1979) along the following lines. Algebraic and differential equations may be ordered by simplicity; the simpler the hypothetical relation between two or more quantities, the greater is its prior probability. If measurement error has a known probability distribution, we can then compute the likelihood of any set of measurement results given an equation relating the measured quantities. It should be clear, then, that with these priors and likelihoods, ratios of posterior probabilities may be computed from measurement results. Jeffreys constructed a Bayesian significance test for the introduction of higher-degree terms

in the equation relating the measured quantities. Roughly, if one's equation fits the data *too* well, then the equation has too many terms and too many arbitrary parameters; and if the equation does not fit the data well enough, then one has not included enough terms and parameters in the equation. The whole business depends, of course, entirely on the ordering of prior probabilities. In his *Theory of Probability* Jeffreys (1967) proposed that the prior probability of a hypothesis decreases as the number of arbitrary parameters increases, but hypotheses having the same number of arbitrary parameters have the same prior probability. This leads immediately to the conclusion that the prior probability of every hypothesis is zero. Earlier, Jeffreys proposed a slightly more complex assignment of priors that did not suffer from this difficulty. The problem is not really one of finding a way to assign finite probabilities to an infinite number of incompatible hypotheses, for there are plenty of ways to do that. The trouble is that it is just very implausible that scientists typically have their prior degrees of belief distributed according to any plausible simplicity ordering, and still less plausible that they would be rational to do so. I can think of very few simple relations between experimentally determined quantities that have with-stood continued investigation, and often simple relations are replaced by relations that are infinitely complex: consider the fate of Kepler's laws. Surely it would be naïve for anyone to suppose that a set of newly measured quantities will truly stand in a simple relation, especially in the absence of a well-confirmed theory of the matter. Jeffreys' strategy requires that we proceed in ignorance of our scientific experience, and that can hardly be a rational requirement (Jeffreys 1973).

Consider another Bayesian attempt, this one due to Mary Hesse. Hesse puts a 'clustering' constraint on prior probabilities: for any positive r, the conjunction of $r+1$ positive instances of a hypothesis is more probable than a conjunction of r positive instances with one negative instance. This postulate, she claims, will lead us to choose, *ceteris, paribus*, the most economical, the simplest, hypotheses compatible with the evidence. Here is the argument:

Consider first evidence consisting of individuals $a_1, a_2, \ldots, a_n$, all of which have properties P and Q. Now consider an individual a_{n+1} with property P. Does a_{n+1} have Q or not? If nothing else is known, the clustering postulate will direct us to predict $Q_{\alpha+1}$ since, *ceteris paribus*, the universe is to be postulated to be as homogeneous as possible consistently with the data... But this is also the prediction that would be made by taking the most economical general law which is both confirmed by the data and of sufficient content to make a prediction about the application of Q to a_{n+1}. For $h =$ 'All P are Q' is certainly more economical than the 'gruified' conflicting hypothesis of equal content h': 'All x up to a_n that are P are Q, and all other x that are P are Q.'

If follows in the [case] considered that if a rule is adopted to choose the prediction resulting from the most probable hypothesis on grounds of content, or, in case of a tie in content, the most economical hypothesis on those of equal content, this rule will yield the same predictions as the clustering postulate.

Here is the argument applied to curve-fitting:

Let f be the assertion that two data points $(x_1, y_1,)$, (x_2, y_2) are obtained from experiments... The two points are consistent with the hypothesis $y = a + bx$,

and also of course with an indefinite number of other hypotheses of the form $y = a_0 + a_1 + \cdots + a_n x_1$, where the values of $a_0, \ldots, a_n$ are not determined by (x_1, y_1), (x_2, y_2). What is the most economical prediction of the y-value of a further point g, where the x-value of g is x_3? Clearly it is the prediction which uses only the information already contained in f, that is, the calculable values of a, b rather than a prediction which assigns arbitrary values to the parameters of a higher-order hypothesis. Hence the most economical prediction is about the point $g = (x_3, a + bx_3)$, which is also the prediction given by the 'simplest' hypothesis on almost all accounts of the simplicity of curves. Translated into probabilistic language, this is to say that to conform to intuitions about economy we should assign higher initial probability to the assertion that points $(x_1, a + bx_1)$, $(x_2, a + bx_2)$, $(x_3, a + bx_3)$ are satisfied by the experiment than to that in which the third point is inexpressible in terms of a and b alone. In this formulation economy is a function of finite descriptive lists of points rather than general hypotheses, and the relevant initial probability is that of a universe containing these particular points rather than that of a universe in which the corresponding general law is true ... Description in terms of a minimum number of parameters may therefore be regarded as another aspect of homogeneity or clustering of the universe. (Hesse 1974: 230–2)

Hesse's clustering postulate applies directly to the curve-fitting case, for her clustering postulate then requires that if two paired values of x and y satisfy the predicate $y = ax + b$, then it is more probable than not that a third pair of values will satisfy the predicate. So the preference for the linear hypothesis in the next instance results from Hesse's clustering postulate and the probability axioms. Unfortunately, with trivial additional assumptions, everything results. For, surely, if $y = a + bx$ is a legitimate predicate, then so is $y = \alpha_1 + b_1 x^2$, for any definite values of a_1 and b_1. Now Hesse's first two data points can be equally well described by $(x_1, a_1 + b_1 x_1^2)$ and $(x_2, a_1 + b_1 x_2^2)$, where

$$b_1 = \frac{y_1 - y_2}{x_1^2 - x_2^2} \quad a_1 = y_1 - x_1^2 \left(\frac{y_1 - y_2}{x_1^2 - x_2^2} \right),$$

Hence her first two data points satisfy both the predicate $y = a + bx$ and the predicate $y = a_1 + b_1 x^2$. So, by the clustering postulate, the probability that the third point satisfies the quadratic expression must be greater than one-half, and the probability that the third point satisfies the linear expression must also be greater than one-half, which is impossible.

Another Bayesian account of our preference for simple theories has recently been offered by Roger Rosencrantz (1976). Suppose that we have some criterion for 'goodness of fit' of a hypothesis to data—for example, confidence regions based on the χ^2 distribution for categorical data, or in curve-fitting perhaps that the average sum of squared deviations is less than some figure. Where the number of possible outcomes is finite, we can compare the number of such possible outcomes that meet the goodness-of-fit criterion with the number that do not. This ratio Rosencrantz calls the 'observed sample coverage' of the hypothesis. Where the possible outcomes are infinite, if the region of possible outcomes meeting the

goodness-of-fit criterion is always bounded for all relevant hypotheses, we can compare the volumes of such regions for different hypotheses, and thus obtain a measure of comparative sample coverage.

It seems plausible enough that the smaller the observed sample coverage of a hypothesis, the more severely it is tested by observing outcomes. Rosencrantz's first proposal is this: the smaller the observed sample coverage, the simpler the hypothesis. But further, he proves the following for hypotheses about categorical data: if H_1 and H_2 are hypotheses with parameters, and H_1 is a special case of H_2 obtained by letting a free parameter in H_2 take its maximum likelihood value, then if we average the likelihood of getting evidence that fits each hypothesis well enough over all the possible parameter values, the average likelihood of H_1 will be greater than the average likelihood of H_2. The conclusion Rosencrantz suggests is that the simpler the theory, the greater the average likelihood of data that fit it sufficiently well. Hence, even if a simple theory has a lower prior probability than more complex theories, because the average likelihood is higher for the simple theory, its posterior probability will increase more rapidly than that of more complex theories. When sufficient evidence has accumulated, the simple theory will be preferred. Rosencrantz proposes to identify average likelihood with support.

Rosencrantz's approach has many virtues; I shall concentrate on its vices. First, observed sample coverage does not correlate neatly with simplicity. If H is a hypothesis, T another utterly irrelevant to H and to the phenomena about which H makes predictions, then H & T will have the same observed sample coverage as does H. Further, if H^* is a de-Occamization of H, then H^* and H will have the same observed sample coverage. Second, Rosencrantz's theorem does not establish nearly enough. It does not establish, for example, that in curve-fitting the average likelihood of a linear hypothesis is greater than the average likelihood of a quadratic or higher-degree hypothesis. We cannot explicate support in terms of average likelihood unless we are willing to allow that evidence supports a de-Occamized hypothesis as much as un-de-Occamized ones, and a hypothesis with tacked-on parts as much as one without such superfluous parts.

Finally, we come to the question of the relevance of evidence to theory. When does a piece of evidence confirm a hypothesis according to the Bayesian scheme of things? The natural answer is that it does so when the posterior probability of the hypothesis is greater than its prior probability, that is, if the conditional probability of the hypothesis on the evidence is greater than the probability of the hypothesis. That is what the condition of positive relevance requires, and that condition is the one most commonly advanced by philosophical Bayesians. The picture is a kinematic one: a Bayesian agent moves along in time having at each moment a coherent set of degrees of belief; at discrete intervals he learns new facts, and each time he learns a new fact, e, he revises his degrees of belief by conditionalizing on e. The discovery that e is the case has confirmed those hypotheses whose probability after the discovery is higher than their probability before. For several reasons, I think this account is unsatisfactory; moreover, I doubt that its difficulties are remediable without considerable changes in the theory.

The first difficulty is a familiar one. Let us suppose that we can divide the consequences of a theory into sentences consisting of reports of actual or possible observations, and simple generalizations of such observations, on the one hand; and on the other hand, sentences that are theoretical. Then the collection of 'observational' consequences of the theory will always be at least as probable as the theory itself; generally, the theory will be less probable than its observational consequences. A theory is never any better established than is the collection of its observational consequences. Why, then, should we entertain theories at all? On the probabilist view, it seems, they are a gratuitous risk. The natural answer is that theories have some special function that their collection of observational consequences cannot serve; the function most frequently suggested is explanation—theories explain; their collection of observational consequences do not. But however sage this suggestion may be, it only makes more vivid the difficulty of the Bayesian way of seeing things. For whatever explanatory power may be, we should certainly expect that goodness of explanation will go hand in hand with warrant for belief; yet, if theories explain, and their observational consequences do not, the Bayesian must deny the linkage. The difficulty has to do both with the assumption that rational degrees of belief are generated by probability measures and with the Bayesian account of evidential relevance. Making degrees of belief probability measures in the Bayesian way already guarantees that a theory can be no more credible than any collection of its consequences. The Bayesian account of confirmation makes it impossible for a piece of evidence to give us more total credence in a theory than in its observational consequences. The Bayesian way of setting things up is a natural one, but it is not inevitable, and wherever a distinction between theory and evidence is plausible, it leads to trouble.

A second difficulty has to do with how praise and blame are distributed among the hypotheses of a theory. Recall the case of Kepler's laws (discussed in Glymour 1981, ch. 2). It seems that observations of a single planet (and, of course, the sun) might provide evidence for or against Kepler's first law (all planets move on ellipses) and for or against Kepler's second law (all planets move according to the area rule), but no observations of a single planet would constitute evidence for or against Kepler's third law (for any two planets, the ratio of their periods equals the $^{3}/_{2}$ power of the ratio of their distances). Earlier [in Ch. 2 of Glymour's *Theory and Evidence*] we saw that hypothetico-deductive accounts of confirmation have great difficulty explaining this elementary judgement. Can the Bayesians do any better? One thing that Bayesians can say (and some have said) is that our degrees of belief are distributed—and historically were distributed—so that conditionalizing on evidence about one planet may change our degrees of belief in the first and second laws, but not our degree of belief in the third law.[5] I don't see that this is an explanation for our intuition at all; on the contrary, it seems merely to restate (with some additional claims) what it is that we want to be explained. Are there any reasons why people had their degrees of belief so distributed? If their beliefs had

[5]This is the account suggested by Horwich (1978).

been different, would it have been equally rational for them to view observations of Mars as a test of the third law, but not of the first? It seems to me that we never succeed in explaining a widely shared judgement about the relevance or irrelevance of some piece of evidence merely by asserting that degrees of belief happened to be so distributed as to generate those judgements according to the Bayesian scheme. Bayesians may instead try to explain the case by appeal to some structural difference among the hypotheses; the only gadget that appears to be available is the likelihood of the evidence about a single planet on various combinations of hypotheses. If it is supposed that the observations are such that Kepler's first and second laws entail their description, but Kepler's third law does not, then it follows that the likelihood of the evidence on the first and second laws—that is, the conditional probability of the evidence given those hypotheses—is unity, but the likelihood of the evidence on the third law may be less than unity. But any attempt to found an account of the case on these facts alone is simply an attempt at a hypothetico-deductive account. The problem is reduced to one already unsolved. What is needed to provide a genuine Bayesian explanation of the case in question (as well as of many others that could be adduced) is a *general* principle restricting conditional probabilities and having the effect that the distinctions about the bearing of evidence that have been noted here do result. Presumably, any such principles will have to make use of relations of content or structure between evidence and hypothesis. The case does nothing to establish that no such principles exist; it does, I believe, make it plain that without them the Bayesian scheme does not *explain* even very elementary features of the bearing of evidence on theory.

A third difficulty has to do with Bayesian kinematics. Scientists commonly argue for their theories from evidence known long before the theories were introduced. Copernicus argued for his theory using observations made over the course of millennia, not on the basis of any startling new predictions derived from the theory, and presumably it was on the basis of such arguments that he won the adherence of his early disciples. Newton argued for universal gravitation using Kepler's second and third laws, established before the *Principia* was published. The argument that Einstein gave in 1915 for his gravitational field equations was that they explained the anomalous advance of the perihelion of Mercury, established more than half a century earlier. Other physicists found the argument enormously forceful, and it is a fair conjecture that without it the British would not have mounted the famous eclipse expedition of 1919. Old evidence can in fact confirm new theory, but according to Bayesian kinematics, it cannot. For let us suppose that evidence e is known before theory T is introduced at time t. Because e is known at t, $\text{prob}_t(e) = 1$. Further, because $\text{prob}_t(e) = 1$, the likelihood of e given T, $\text{prob}_t(e, T)$, is also 1. We then have

$$\text{prob}_t(T,\ e) = \frac{\text{prob}_t(T) \times \text{prob}_t(e,\ T)}{\text{prob}_t(e)} = \text{prob}_t(T).$$

The conditional probability of T on e is therefore the same as the prior probability of T: e cannot constitute evidence for T in virtue of the positive relevance condition

nor in virtue of the likelihood of e on T. None of the Bayesian mechanisms apply, and if we are strictly limited to them, we have the absurdity that old evidence cannot confirm new theory. The result is fairly stable. If the probability of e is very high but not unity, $\text{prob}_t(e, T)$ will still be unity if T entails e, and so $\text{prob}_r(T, e)$ will be very close to $\text{prob}_t(T)$. How might Bayesians deal with the old evidence/new theory problem?[6] Red herrings abound. The prior probability of the evidence, Bayesians may object, is not really unity; when the evidence is stated as measured or observed values, the theory does not really entail that those exact values obtain; an ideal Bayesian would never suffer the embarrassment of a novel theory. None of these replies will do: the acceptance of old evidence may make the degree of belief in it as close to unity as our degree of belief in some bit of evidence ever is; although the exact measured value (of, e.g., the perihelion advance) may not be entailed by the theory and known initial conditions, that the value of the measured quantity lies in a certain interval may very well be entailed, and that is what is believed anyway; and, finally, it is beside the point that an ideal Bayesian would never face a novel theory, for the idea of Bayesian confirmation theory is to explain scientific inference and argument by means of the assumption that good scientists are, about science at least, approximately ideal Bayesians, and we have before us a feature of scientific argument that seems incompatible with that assumption.

A natural line of defence lies through the introduction of counterfactual degrees of belief. When using Bayes's rule to determine the posterior probability of a new theory on old evidence, one ought not to use one's actual degree of belief in the old evidence, which is unity or nearly so; one ought instead to use the degree of belief one would have had in e if . . . The problem is to fill in the blanks in such a way that it is both plausible that we have the needed counterfactual degrees of belief, and that they do serve to determine how old evidence bears on new theory. I tend to doubt that there is such a completion. We cannot merely throw e and whatever entails e out of the body of accepted beliefs; we need some rule for determining a counterfactual degree of belief in e and a counterfactual likelihood of e on T. To simplify, let us suppose that T does logically entail e, so that the likelihood is fixed.

If one flips a coin three times and it turns up heads twice and tails once, in using this evidence to confirm hypotheses (e.g. of the fairness of the coin), one does not take the probability of two heads and one tail to be what it is after the flipping—namely, unity—but what it was before the flipping. In this case there is an immediate and natural counterfactual degree of belief that is used in conditionalizing by Bayes's rule. The trouble with the scientific cases is that no such immediate and natural alternative distribution of degree of belief is available. Consider someone trying, in a Bayesian way, to determine in 1915 how much Einstein's derivation

[6] All of the defences sketched below were suggested to me by one or another philosopher sympathetic to the Bayesian view; I have not attributed the arguments to anyone for fear of misrepresenting them. None the less, I thank Jon Dorling, Paul Teller, Daniel Garber, Ian Hacking, Patrick Suppes, Richard Jeffrey, and Roger Rosencrantz for valuable discussions and correspondence on the point at issue.

of the perihelion advance confirmed general relativity. There is no single event, like the coin flipping, that makes the perihelion anomaly virtually certain. Rather, Leverrier first computed the anomaly in the middle of the nineteenth-century; Simon Newcomb calculated it again around 1890, using Leverrier's method but new values for planetary masses, and obtained a substantially higher value than had Leverrier. Both Newcomb and Leverrier had, in their calculations, approximated an infinite series by its first terms without any proof of convergence, thus leaving open the possibility that the entire anomaly was the result of a mathematical error. In 1912 Eric Doolittle calculated the anomaly by a wholly different method, free of any such assumption, and obtained virtually the same value as had Newcomb.[7] For actual historical cases, unlike the coin-flipping case, there is no single counterfactual degree of belief in the evidence ready to hand, for belief in the evidence sentence may have grown gradually—in some cases, it may have even waxed, waned, and waxed again. So the old evidence/new theory problem cannot be assimilated to coin flipping.

The suggestion that what is required is a counterfactual degree of belief is tempting, none the less; but there are other problems with it besides the absence of any unique historical degree of belief. A chief one is that various ways of manufacturing counterfactual degrees of belief in the evidence threaten us with incoherence. One suggestion, for example, is the following, used implicitly by some Bayesian writers. At about the time T is introduced, there will be a number of alternative competing theories available; call them $T_1, T_2, \ldots, T_k$, and suppose that they are mutually exclusive of T and of each other. Then $P(e)$ is equal to

$$P(T_1)\,P(e, T_1) + P(T_2)\,P(e, T_2) + \cdots + P(T_k)\,P(e, T_k) + P(\sim(T_1 \vee \cdots \vee T_k) \times P(e, \sim T_1 \vee \cdots \vee T_k)),$$

and we may try to use this formula to evaluate the counterfactual degree of belief in e. The problem is with the last term. Of course, one could suggest that this term just be ignored when evaluating $P(e)$, but it is difficult to see within a Bayesian framework any rationale at all for doing so. For if one does ignore this term, then the collection of prior probabilities used to evaluate the posterior probability of T will not be coherent unless either the likelihood of e on T is zero or the prior probability of T is zero. One could remedy this objection by replacing the last term by

$$P(T)P(e, T),$$

but this will not do either, for if one's degree of belief in

$$P(T_1 \vee T_2 \vee \cdots \vee T_k \vee T)$$

[7]The actual history is still more complicated. Newcomb and Doolittle obtained values for the anomaly differing by about 2 s of are per century. Early in the 1920s. Grossmann discovered that Newcomb had made an error in calculation of about that magnitude.

is not unity, then the set of prior degrees of belief will still be incoherent. Moreover, not only will it be the case that if the actual degree of belief in e is replaced by a counterfactual degree of belief in e according to either of these proposals, then the resulting set of priors will be incoherent, it will further be the case that if we conditionalize on e the resulting conditional probabilities will be incoherent. For example, if we simply delete the last term, one readily calculates that

$$P\,(T_1 \vee \cdots \vee T_k,\ e) = \frac{P\,(T_1 \vee \cdots \vee T_k)\ P\,(e,\ T_1 \vee \cdots \vee T_k)}{P\,(e,\ T_1 \vee \cdots \vee T_k)\ P\,(T_1 \vee \cdots \vee T_k)} = 1,$$

and further that

$$P\,(T,\ e) = \frac{P(T)\ P\,(e,\ T)}{P\,(e,\ T_1 \vee \cdots \vee T_k)\ P\,(T_1 \vee \cdots \vee T_k)}.$$

But because T is supposed inconsistent with $T_1 \vee \cdots \vee T_k$ and $P(T, e)$ is not zero, this is incoherent.

Let us return to the proposal that when new theory confronts old evidence, we should look backwards to the time when the old evidence e had not yet been established and use for the prior probability of e whatever degree of belief we would have had at that time. We cannot just stick in such a counterfactual value for the prior probability of e and change nothing else without, as before, often making both prior and conditionalized probabilities incoherent. If we give all of our sentences the degree of belief they would have had in the relevant historical period (supposing we somehow know what period that is) and then conditionalize on e, incoherence presumably will not arise; but it is not at all clear how to combine the resulting completely counterfactual conditional probabilities with our actual degrees of belief. It does seem to me that the following rather elaborate procedure will work when a new theory is introduced. Starting with your actual degree of belief function P, consider the degree of belief you would have had in e in the relevant historical period, call it $H(e)$. Now change P by regarding $H(e)$ as an arbitrary change in degree of belief in e and using Richard Jeffrey's (1965) rule,

$$P'(S) = H(e)P\,(S, e) + (1 - H(e))\,P\,(S, \sim e)\,.$$

Jeffrey's rule guarantees that P' is a probability function. Finally, conditionalize on e:

$$P''(S) = P'\,(S, e)\,,$$

and let P'' be your new actual degree of belief function. (Alternatively, P'' can be formed by using Jeffrey's rule a second time.)

There remain a number of objections to the historical proposal. It is not obvious that there are, for each of us, degrees of belief we personally would have had in some historical period. It is not at all clear which historical period is the relevant

one. Suppose, for example, that the gravitational deflection of sunlight had been determined experimentally around 1900, well before the introduction of general relativity.[8] In trying to assess the confirmation of general relativity, how far back in time should a twentieth-century physicist go under this supposition? If only to the nineteenth, then if he would have shared the theoretical prejudices of the period, gravitational deflection of light would have seemed quite probable. Where ought he to stop, and why? But laying aside these difficulties, it is implausible indeed that such a historical Bayesianism, however intriguing a proposal, is an accurate account of the principles by which scientific judgements of confirmation are made. For if it were, then we should have to condemn a great mass of scientific judgements on the grounds that those making them had not studied the history of science with sufficient closeness to make a judgement as to what their degrees of belief would have been in relevant historical periods. Combined with the delicacy that is required to make counterfactual degrees of belief fit coherently with actual ones, these considerations make me doubt that we should look to counterfactual degrees of belief for a plausible Bayesian account of how old evidence bears on new theory.

Finally, consider a quite different Bayesian response to the old evidence/new theory problem. Whereas the ideal Bayesian agent is a perfect logician, none of us are, and there are always consequences of our hypotheses that we do not know to be consequences. In the situation in which old evidence is taken to confirm a new theory, it may be argued that there is *something* new that is learned, and typically, what is learned is that the old evidence is entailed by the new theory. Some old anomalous result is lying about, and it is not this old result that confirms a new theory, but rather the new discovery that the new theory entails (and thus explains) the old anomaly. If we suppose that semi-rational agents have degrees of belief about the entailment relations among sentences in their language, and that

$$P\left(h \middle| -e\right) = 1 \;\; \textit{implies} \;\; P(e, h) = 1,$$

this makes a certain amount of sense. We imagine the semi-rational Bayesian changing his degree of belief in hypothesis h in light of his new discovery that h entails e by moving from his prior degree of belief in h to his conditional degree of belief in h given that e, that $h \vdash e$, and whatever background beliefs there may be. Old evidence can, in this vicarious way, confirm a new theory, then, provided that

[8] Around 1900 is fanciful, before general relativity is not. In 1914 E. Freundlich mounted an expedition to Russia to photograph the eclipse of that year in order to determine the gravitational deflection of starlight. At that time, Einstein had predicted an angular deflection for light passing near the limb of the sun that was equal in value to that derived from Newtonian principles by Soldner in 1801. Einstein did not obtain the field equations that imply a value for the deflection equal to twice the Newtonian value until late in 1915. Freundlich was caught in Russia by the outbreak of World War I, and was interned there. Measurement of the deflection had to wait until 1919.

$$P\left(h, b\&e\&\left(h\middle|-e\right)\right) > P\left(h, b\&e\right).$$

Now, in a sense, I believe this solution to the old evidence/new theory problem to be the correct one; what matters is the discovery of a certain logical or structural connection between a piece of evidence and a piece of theory, and it is in virtue of that connection that the evidence, if believed to be true, is thought to be evidence for the bit of theory. What I do not believe is that the relation that matters is simply the entailment relation between the theory, on the one hand, and the evidence, on the other. The reasons that the relation cannot be simply that of entailment are exactly the reasons why the hypothetico-deductive account (see Glymour 1981, ch. 2) is inaccurate; but the suggestion is at least correct in sensing that our judgement of the relevance of evidence to theory depends on the perception of a structural connection between the two, and that degree of belief is, at best, epiphenomenal. In the determination of the bearing of evidence on theory, there seem to be mechanisms and stratagems that have no apparent connection with degrees of belief, which are shared alike by people advocating different theories. Save for the most radical innovations, scientists seem to be in close agreement regarding what would or would not be evidence relevant to a novel theory; claims as to the relevance to some hypothesis of some observation or experiment are frequently buttressed by detailed calculations and arguments. All of these features of the determination of evidential relevance suggest that that relation depends somehow on structural, objective features connecting statements of evidence and statements of theory. But if that is correct, what is really important and really interesting is what these structural features may be. The condition of positive relevance, even if it were correct, would simply be the least interesting part of what makes evidence relevant to theory.

None of these arguments is decisive against the Bayesian scheme of things, nor should they be; for in important respects that scheme is undoubtedly correct. But taken together, I think they do at least strongly suggest that there must be relations between evidence and hypotheses that are important to scientific argument and to confirmation but to which the Bayesian scheme has not yet penetrated.

References

Carnap, R. (1950). *The logical foundations of probability*. Chicago: University of Chicago Press.
Glymour, C. (1981). *Theory and evidence*. Chicago: University of Chicago Press.
Hesse, M. (1974). *The structure of scientific inference*. Berkeley: University of California Press.
Horwich, P. (1978). An appraisal of Glymour's confirmation theory. *Journal of Philosophy, 75*, 98–113.
Jeffrey, R. (1965). *The logic of decision*. New York: McGraw-Hill.
Jeffreys, H. (1967). *Theory of probability*. Oxford: Clarendon.
Jeffreys, H. (1973). *Scientific inference*. Cambridge: Cambridge University Press.
Kyburg, H. (1978). Subjective probability: Criticisms, reflections and problems. *Journal of Philosophical Logic, 7*, 157–180.

Putnam, H. (1967). Probability and confirmation. In S. Morgenbesser (Ed.), *Philosophy of science today*. New York: Basic Books.

Rosencrantz, R. (1976). Simplicity. In W. Harper & C. Hooker (Eds.), *Foundations and philosophy of statistical inference*. Boston: Reidel.

Salmon, W. C. (1969). *Foundations of scientific inference*. Pittsburgh: University of Pittsburgh Press.

Savage, L. (1972). *The foundations of statistics*. New York: Dover.

Shimony, A. (1970). Scientific inference. In R. G. Colodny (Ed.), *The nature and function of scientific theories* (pp. 79–179). Pittsburgh: University of Pittsburgh Press.

Chapter 9
Discussion: A Mistake in Dynamic Coherence Arguments?

Brian Skyrms

Static Coherence of Degrees of Belief

The person whose degrees of belief are being tested for coherence acts as a bookie. She posts her fair prices for wagers corresponding to her degrees of belief. Her degrees of belief are *incoherent* if a cunning bettor can make a Dutch book against her with a finite system of wagers—that is, there is a finite set of wagers individually perceived as fair, whose net payoff is a loss in every possible future. Otherwise her degrees of belief are *coherent*. De Finetti ([1937] 1980) proved the following theorem: *Degrees of belief are coherent if and only if they are finitely additive probabilities*.

Obviously, if a Dutch book can be made with a finite number of fair transactions, it can be made with a finite number of uniformly favorable transactions. The bettor pays some small transaction premium ε to the bookie for each of the n transactions where $n\varepsilon$ is less than the guaranteed profit that the bettor gets under the Dutch book based on fair prices. Let us bear in mind that this point applies equally well in what follows.

Received April 1992; revised June 1992.
Send reprint requests to the author, Department of Philosophy, University of California, Irvine, CA 92717, USA.

B. Skyrms (✉)
Department of Philosophy, University of California, Irvine, CA, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_9

Dynamic Coherence for Updating Rules

The epistemologist acts as bookie. Her updating rule is public knowledge. Today she posts her fair prices, and does business. Tomorrow she makes an observation (with a finite number of possible outcomes each of which has positive prior probability) and updates her fair prices according to her updating rule. The updating rule is thus a *function* from possible observations to revised fair prices. The day after tomorrow she posts prices again, and does business. The pair consisting of the (1) her fair prices for today and (2) her updating function will be called the bookie's *epistemic strategy*.

The bookie's *epistemic strategy* is *coherent* if *there is no* possible bettor's strategy which makes a Dutch book against him (the bettor's strategy being a pair consisting of (1) a finite number of transactions today at the bookie's posted prices and (2) a function taking possible observations into a finite number of transactions the day after tomorrow at the prices that the bookie will post according to her epistemic strategy). Lewis (reported in Teller 1973) proves that *the epistemologist's strategy is coherent only if her degrees of belief today are finitely additive probabilities and her updating rule is Bayes's rule of conditioning*. The "only if" can be strengthened to "if and only if" (see section "The converse"). (For generalizations of this theorem see van Fraassen 1984 and Skyrms 1987, 1990.)

Notice that the relevant notions of coherence and incoherence here apply not just to the *pair* of degrees of belief for today and the day after tomorrow, but rather to an *epistemic strategy*, which is a more complicated object. A focus on the former notion leads understandably to skepticism regarding dynamic coherence, as in Hacking (1967), Kyburg (1978), and Christensen (1991).

The Dynamic Dutch Book

Coherence of degrees of belief today is the static case. It remains to show that for any non-Bayes updating rule, there is a bettor's strategy which makes a Dutch book. Let the conditional probability of A on e, that is $\Pr(A \,\&\, e)/\Pr(e)$, be symbolized as usual, as $\Pr(A|e)$, and let the probability that the updating rule gives A if e is observed be $\Pr_e(A)$. If the predictor's rule disagrees with conditioning, then for some possible evidential result e and some A, $\Pr_e(A)$ is not $\Pr(A|e)$. Suppose that $\Pr(A|e) > \Pr_e(A)$. (The other case is similar.) Let the discrepancy be $\delta = \Pr(A|e) - \Pr_e(A)$. *Here is a bettor's strategy which makes a Dutch book:*

TODAY: Offer to sell the bookie at her fair price:

1: [\$1 if A & e, 0 otherwise]
2: [\$$\Pr(A|e)$ if not-e, 0 otherwise]
3: [\$$\delta$ if e, 0 otherwise]

DAY AFTER TOMORROW:
If e was observed, offer to buy [\$1 if A, 0 otherwise] for its current fair price, $\Pr_e(A) = \Pr(A|e) - \delta$.

Then in every possible situation, the bookie loses $\$\delta \Pr(e)$.

The Converse

If the bookie has the strategy of updating by Bayes's rule of conditioning, then every payoff that a bettor's strategy can achieve can be achieved by betting only today (see Skyrms 1987). This reduces our case to the static case. Thus, by de Finetti's result, if the epistemologist's prior degrees of belief are finitely additive probabilities and her updating rule is Bayes's rules of conditioning, then she is dynamically coherent.

Sequential Analysis 1: A Mistake in the Dynamic Coherence Argument?

Maher's (1992b) objection is that the bookie will see it coming and refuse to bet. This is made precise by modeling the bookie's situation as a sequential choice problem, as shown in Fig. 9.1. The bookie sees that if she bets today and e occurs, then at decision node 2, she will find the cunning bettor's offer fair according to her revised probability, $\Pr_e(A)$. Thus she sees that betting today leads to a sure loss. Since she prefers net gain of zero to a sure loss, she refuses to bet today—frustrating the cunning bettor who goes home unable to execute his plan.

The first thing that must be said about "Maher's objection" is that it is misleading to represent it as showing a "mistake" in the dynamic coherence theorem. Under the conditions of the theorem the bookie posts her fair prices for today and honors them. There is no provision for changing one's mind when approached by a cunning bettor who discloses his strategy, nor indeed any mention of a requirement that the cunning bettor disclose his strategy prior to the initial transaction. But Maher might be read as suggesting a different conception of dynamic coherence in this setting:

> The epistemologist acts as bookie. Her updating rule is public knowledge. Today she posts her tentative fair prices, but in fact does business only with bettors who disclose their strategies in advance, and does so on the basis of sequential decision analysis. Tomorrow she makes an observation (with a finite number of possible outcomes each of which has positive prior probability) and updates her probabilities according to her updating rule. The day after tomorrow she posts prices again, and does business according to those prices.
>
> She is *coherent* if *there is no* possible bettor's strategy which makes a Dutch book against her.

This is an interesting modification of the usual notion of dynamic coherence, and it merits investigation. Is it a better motivated conception of dynamic coherence? What differences does it make?

Sequential Analysis 2: A Mistake in the Mistake?

A natural reaction to Maher's line might be to say that the redefinition unfairly prejudices the case against dynamic coherence arguments. It is therefore of some interest to see that the dynamic Dutch book still goes through under the revised scenario.

There is a gratuitous assumption in the analysis presented in Fig. 9.1. Why is it assumed that the cunning bettor will just go home if the bookie refuses to bet today? The bettor's strategy which I presented says otherwise. The bettor will make an offer the day after tomorrow if *e* was observed. So the branch of the decision tree where the bookie refuses transactions today cannot simply be assumed to have payoff of zero, but requires further analysis. This is done in Fig. 9.2.

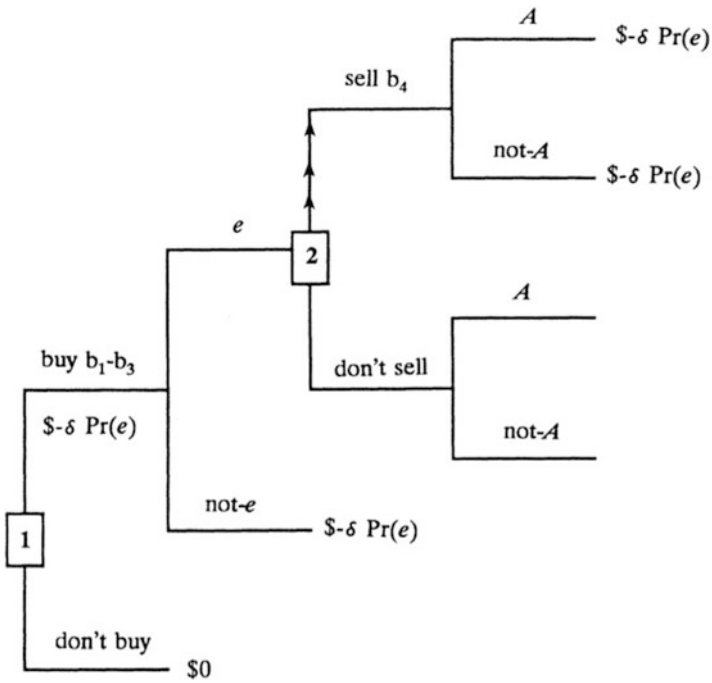


Fig. 9.1 Maher-Levi sequential analysis

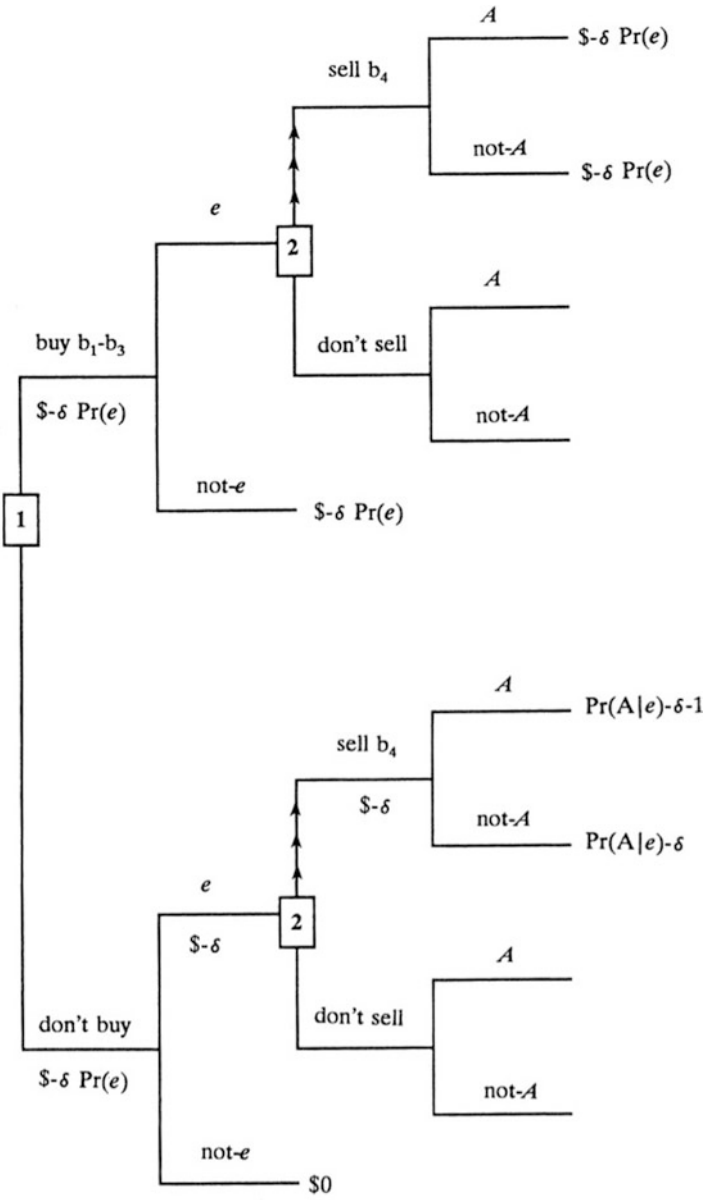


Fig. 9.2 Full sequential analysis

Note that the bookie knows that if e is observed, she will accept the offer the day after tomorrow for the same reason on the lower path as on the upper. Deciding now not to bet ever is not an option. If the offer the day after tomorrow is accepted but the offer today was not and e and A both happen, then the net payoff is the price the cunning bettor paid, $\$\Pr(A|e) - \delta$, less the lost bet, $\$ - 1$, as shown. If e occurs but A does not, the net payoff is just $\$\Pr(A|e) - \delta$. For the bookie's current analysis of this decision tree, to get the relevant expectation over A occurring or not we average using as weights her current conditional probabilities, $\Pr(A|e)$ and $\Pr(-A|e)$. Thus the value at the node where the bookie refused to bet today and where e is observed tomorrow is

$$\Pr\left(A\middle|e\right)\$\left[\left\{\Pr\left(A\middle|e\right)-\delta\right\}-1\right]+\left[1-\Pr\left(A\middle|e\right)\right]\$\left[\Pr\left(A\middle|e\right)-\delta\right]=\$-\delta.$$

Then the value at the node where the bookie refused to bet today is not 0 but rather $\$ - \delta \Pr(e)$. This is just the same as the value at the node where the bookie agrees to bet today.

In fact, if we consider the version of the Dutch-book strategy where the bettor adds an ε premium for each transaction, the upper branch involves four transactions and the lower branch involves only one, so the upper branch has a higher payoff than the lower branch. *Even though the bookie sees it coming, she will prefer the sure loss of the upper branch because doing so looks strictly better to her than the alternative*.

Sequential Analysis 3: What Makes the Cunning Bettor Tick?

Why did the cunning bettor adopt a strategy of staying around if the bookie decided not to bet today? The official answer in sections "Dynamic coherence for updating rules" and "The dynamic Dutch book" is "Don't ask". *Any* bettor's strategy which makes a Dutch book will prove incoherence. But, as Levi (1991) points out, that sort of analysis proceeds in strategic normal form rather than in extensive form. Might it be that the cunning bettor's strategy described would have to be *sequentially irrational*? That is to say, might it not be that staying around and betting the day after tomorrow if the bookie decided not to bet today would not maximize expected utility for the cunning bettor in the belief state he would have in that case the day after tomorrow? If this could be shown, then the cunning bettor's strategy that I have described would have to rest on a noncredible threat, and the significance of the analysis of the previous section would be called into question. (For discussion of such noncredible threats in extensive form games and of sequential rationality, see Selten 1975 and Kreps and Wilson 1982.)

But such is not the case. Suppose that the bettor is a Bayesian; that he starts out with exactly the same degrees of belief as the bookie; and that he updates by conditioning. If e is observed tomorrow—whether or not the bookie accepted the

bet today—he conditions on e and the day after tomorrow his fair price for $b4$ is \$ $\mathrm{pr}(A|e)$. But his strategy only commits him to offering to pay the bookie's fair price, \$ $\mathrm{pr}(A|e) - \delta$, to buy back $b4$ for what he perceives as a net gain in expected utility of \$$\delta$. This bettor's threat to stick around and bet the day after tomorrow if e, even if the bookie declines to bet today, is perfectly credible and consistent with sequential rationality. If he is called upon to carry out the threat, he maximizes expected utility by doing so.

Strategic Rationality

Let us explicitly model the bookie's choice of an updating strategy. The bookie and the bettor start out with identical priors. The bettor updates by conditioning. First the bookie chooses an updating strategy. Then the bettor bets, the evidence comes in, the bookie updates according to her updating rule, and the bettor bets again. The bookie's initial strategy is either to choose updating by conditioning or not.

If the bookie chooses the strategy of updating by conditioning, then the fair prices of the bookie and bettor agree at all times. Thus either no transactions are made, or any transactions have net change in expected utility of 0 for both players. The bookie's expected utility of choosing the strategy of updating by conditioning is zero. If, however, the bookie chooses an updating strategy at variance with conditioning then, for the bettor, the expected utility of betting is greater than that of not betting (section "Sequential analysis 3: what makes the cunning bettor tick?") and the net expected utility for the bookie is negative (section "Sequential analysis 2: a mistake in the mistake?"). At the first choice point the bookie is strictly better off by choosing the rule of updating by conditioning.

Thus the strategy combination in which the bookie updates by conditioning and the bettor does not bet at all is an *equilibrium* in the sense that no player will perceive it in his or her interest at any decision node to deviate from that strategy. But no strategy combination in which the bookie chooses a strategy at variance with conditioning is such an equilibrium.

The Bottom Line

Two ways of strengthening the requirements for a dynamic Dutch book were suggested by the discussions of Levi (1987) and Maher: (1) We require the cunning bettor to disclose his strategy, and allow the bookie to use knowledge of that strategy in a sequential analysis when deciding whether to bet today or not, and (2) we require that the cunning bettor's strategy itself be sequentially rational. The somewhat surprising result is that the additional restrictions made no difference. The bookie whose epistemic strategy is at odds with conditioning is also subject to a Dutch book in this stronger sense. "Seeing it coming" does not help. It is at

the very least a noteworthy property of the rule of conditioning that in this sort of epistemic situation, it alone is immune from a Dutch book under either the original or strengthened requirements.

Postscript: Conditioning, Coherence and Rationality

Many of the concerns of Levi and Maher have not been addressed in the foregoing. Levi is concerned to resist the doctrine of "confirmational tenacity", according to which the only legitimate way in which to update is by conditioning. Maher wishes to resist the doctrine that rationality requires dynamic coherence at all costs. Does the foregoing show that conditioning is the only coherent way to ever update one's probabilities? Does it show that rationality requires coherence at all costs?

I agree with Levi and Maher in answering "no" to both questions. With regard to the first, let me emphasize that the Lewis proof takes place within the structure of a very special epistemic model. In that context it shows that the rule of conditioning is the unique dynamically coherent updating rule. It does not show that one must have an updating rule. It does not apply to other epistemic situations which should be modeled differently. The modeling of a variety of epistemic situations and the investigation of varieties of dynamic coherence in such situations is an ongoing enterprise (in which I take it that both Levi and I are engaged; see Skyrms 1990 for further discussion).

Maher is concerned that an uncritical doctrine of "dynamic coherence at all costs" could lead one to crazy belief changes and disastrous actions. Should Ulysses have changed to 1 his prior probability of safe sailing conditional on hearing the Sirens' song so that subsequently his belief change would be in accordance with the rule of conditioning? Nothing in the foregoing implies that he should. In the first place, there is something a little odd in thinking that one achieves dynamic coherence by changing the original prior pr_1 to the revised prior pr_2 so that the change to pr_3 will agree with conditioning. What about the change from pr_1 to pr_2? But, more fundamentally, I would agree with Maher that rationality definitely does not require coherence *at all costs*. Where costs occur they need to be weighed against benefits. There are lucid discussions of this matter in Maher (1992a, b). These things said, it remains that in the Lewis epistemic model under the "sequentialized" notion of dynamic coherence, the unique coherent updating rule is the rule of conditioning.

Acknowledgement I would like to thank Brad Armendt, Ellery Eells, Isaac Levi, Patrick Maher and an anonymous referee for helpful comments on an earlier draft of this note. I believe that Maher, Levi and I are now in substantial agreement on the issues discussed here, although differences in emphasis and terminology may remain.

References

Christensen, D. (1991). Clever bookies and coherent beliefs. *The Philosophical Review, 100*, 229–247.

de Finetti, B. ([1937] 1980). Foresight: Its logical laws, its subjective sources, translated in H. E. Kyburg, Jr., & H. Smokler (Eds.), *Studies in subjective probability* (pp. 93–158). (Originally published as "La Prevision: ses lois logiques, ses sources subjectives", *Annales de l'Institut Henri Poincaré, 7*, 1–68.) Huntington: Kreiger.

Hacking, I. (1967). Slightly more realistic personal probability. *Philosophy of Science, 34*, 311–325.

Kreps, D., & Wilson, R. (1982). Sequential equilibria. *Econometrica, 50*, 863–894.

Kyburg, H. (1978). Subjective probability: Criticisms, reflections and problems. *The Journal of Philosophical Logic, 7*, 157–180.

Levi, I. (1987). The demons of decision. *The Monist, 70*, 193–211.

Levi, I. (1991). Consequentialism and sequential choice. In M. Bacharach & S. Hurley (Eds.), *Foundations of decision theory* (pp. 92–146). Oxford: Basil Blackwell.

Maher, P. (1992a). *Betting on theories*. Cambridge: Cambridge University Press.

Maher, P. (1992b). Diachronic rationality. *Philosophy of Science, 59*, 120–141.

Selten, R. (1975). Reexamination of the perfectness concept of equilibrium in extensive form games. *International Journal of Game Theory, 4*, 25–55.

Skyrms, B. (1987). Dynamic coherence and probability kinematics. *Philosophy of Science, 54*, 1–20.

Skyrms, B. (1990). *The dynamics of rational deliberation*. Cambridge: Harvard University Press.

Teller, P. (1973). Conditionalization and observation. *Synthese, 26*, 218–258.

van Fraassen, B. (1984). Belief and the will. *Journal of Philosophy, 81*, 235–256.

Chapter 10
Some Problems for Conditionalization and Reflection

Frank Arntzenius

I will present five puzzles which show that rational people can update their degrees of belief in manners that violate Bayesian Conditionalization and van Fraassen's Reflection Principle. I will then argue that these violations of Conditionalization and Reflection are due to the fact that there are two, as yet unrecognized, ways in which the degrees of belief of rational people can develop.

Two Roads to Shangri La

Every now and then the guardians to Shangri La will allow a mere mortal to enter that hallowed ground. You have been chosen because you are a fan of the Los Angeles Clippers. But there is an ancient law about entry into Shangri La: you are only allowed to enter, if, once you have entered, you no longer know by what path you entered. Together with the guardians you have devised a plan that satisfies this law. There are two paths to Shangri La, the Path by the Mountains, and the Path by the Sea. A fair coin will be tossed by the guardians to determine which path you will take: if heads you go by the Mountains, if tails you go by the Sea. If you go by the Mountains, nothing strange will happen: while traveling you will see the glorious Mountains, and even after you enter Shangri La you will forever retain your memories of that Magnificent Journey. If you go by the Sea, you will revel in the Beauty of the Misty Ocean. But, just as you enter Shangri La, your memory of this Beauteous Journey will be erased and be replaced by a memory of the Journey by the Mountains.

F. Arntzenius (✉)
University of Oxford, Oxford, UK
e-mail: frank.arntzenius@philosophy.ox.ac.uk

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_10

Suppose that in fact you travel by the Mountains. How will your degrees of belief develop? Before you set out your degree of belief in heads will be ½. Then, as you travel along the Mountains and you gaze upon them, your degree of belief in heads will be 1. But then, once you have arrived, you will revert to having degree of belief ½ in heads. For you will know that you would have had the memories that you have either way, and hence you know that the only relevant information that you have is that the coin was fair.

This seems a bizarre development of degrees of belief. For as you are traveling along the Mountains, you know that your degree of belief in heads is going to go down from 1 to ½. You do not have the least inclination to trust those future degrees of belief. Those future degrees of belief will not arise because you will acquire any evidence, at least not in any straightforward sense of "acquiring evidence". Nonetheless you think you will behave in a fully rational manner when you acquire those future degrees of belief. Moreover, you know that the development of your memories will be completely normal. It is only because something strange would have happened to your memories had the coin landed tails, that you are compelled to change your degrees of belief to ½ when that counterfactual possibility would have occurred.

The Prisoner

You have just been returned to your cell on death row, after your last supper. You are to be executed tomorrow. You have made a last minute appeal to President George W. Bush for clemency. Since Dick Cheney is in hospital and can not be consulted, George W. will decide by flipping a coin: heads you die, tails you live. His decision will be made known to the prison staff before midnight. You are friends with the prison officer that will take over the guard of your cell at midnight. He is not allowed to talk to you, but he will tell you of Bush's decision by switching the light in your cell off at the stroke of midnight if it was heads. He will leave it on if it was tails. Unfortunately you don't have a clock or a watch. All you know is that it is now 6 pm since that is when prisoners are returned to their cells after supper. You start to reminisce and think fondly of your previous career as a Bayesian. You suddenly get excited when you notice that there is going to be something funny about the development of your degrees of belief. Like anybody else, you don't have a perfect internal clock. At the moment you are certain that it is 6 pm, but as time passes your degrees of belief are going to be spread out over a range of times. What rules should such developments satisfy?

Let us start on this problem by focusing on one particularly puzzling feature of such developments. When in fact it is just before midnight, say 11.59 pm, you are going to have a certain, non-zero, degree of belief that it is now later than midnight. Of course, at 11.59 pm the light in your cell is still going to be on. Given that at this time you will have a non-zero degree of belief that it is after midnight, and given that in fact you will see that the light is still on, you will presumably take it

that the light provides some evidence that the outcome was tails. Indeed, it seems clear that as it gets closer to midnight, you will monotonically increase your degree of belief in tails. Moreover you know in advance that this will happen. This seems puzzling. Of course, after midnight, your degree of belief in tails will either keep on increasing, or it will flip to 0 at midnight and stay there after midnight. But that does not diminish the puzzlement about the predictable and inevitable increase in your degree of belief in tails prior to midnight. In fact, it seems that this increase is not merely puzzling, it seems patently irrational. For since this increase is entirely predictable, surely you could be made to lose money in a sequence of bets. At 6 pm you will be willing to accept a bet on heads at even odds, and at 11.59 pm you will, almost certainly, be willing to accept a bet on tails at worse than even odds. And that adds up to a sure loss. And surely that means you are irrational.

Now, one might think that this last argument shows that your degree of belief in tails in fact should not go up prior to midnight. One might indeed claim that since your degree of belief in heads should remain ½ until midnight, you should adjust your idea of what time it is when you see that the light is still on, rather than adjust your degree of belief in tails as time passes. But of course, this suggestion is impossible to carry out. Armed with an imperfect internal clock, you simply can not make sure that your degree of belief in heads stays ½ until midnight, while allowing it to go down after midnight. So how should they develop?

Let us start with a much simpler case. Let us suppose that there is no coin toss and no light switching (and that you know this). You go into you cell at 6 pm. As time goes by there will be some development of your degrees of belief as to what time it is. Let us suppose that your degrees of belief in possible times develop as pictured in the top half of Fig. 10.1.

Next, let us ask how your degrees of belief should develop were you to know with certainty that the guard will switch the light off at 12 pm. It should be clear that then at 11.59 pm your degree of belief distribution should be entirely confined

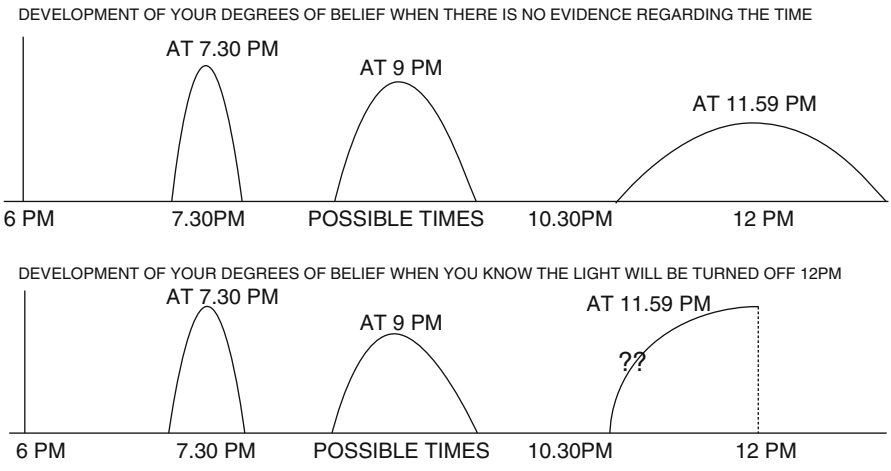


Fig. 10.1 Prisoner without evidence

to the left of 12 pm, as depicted in the bottom half of Fig. 10.1. For at 11.59 pm the light will still be on, so that you know that it must be before 12 pm. But other than that it should be entirely confined to the left of 12 pm, it is not immediately clear exactly what you degree of belief distribution should be at 11.59 pm. It is not even obvious that there should be a unique answer to this question. However, a very simple consideration leads to a unique answer.

Suppose that, even though the guard is going to switch the light off at 12 pm, you were not told that the guard is going to switch the light off at 12 pm. Then the development of your degrees of belief would be as pictured in the top half of Fig. 10.1. Next, suppose that at 11.59 pm your are told that the guard will switch the light off at 12 pm, but you are not told that it is now 11.59 pm. Obviously, since the light is still on you can infer that it is prior to 12 pm. Surely you should update your degrees of belief by conditionalization: you should erase that part of your degree of belief distribution that is to the right of 12 pm, and re-normalize the remaining part (increase the remaining part proportionally). Now it is clear that this is also the degree of belief distribution that you should have arrived at had you known all along that the guard would turn the light off at 12 pm. For either way you have accumulated exactly the same relevant information and experience by 11.59 pm. This uniquely determines how your degree of belief distribution should develop when you know all along that the guard will turn the light off at 12 pm. At any time this (constrained) distribution should be the distribution that you arrive at by conditionalizing the distribution that you have if you have no evidence regarding the time, on the fact that it is now before 12 pm. One can picture this development in the following way. One takes the development of the top part of Fig. 10.1. As this distribution starts to pass through the 12 pm boundary, the part that passes through this boundary gets erased, and, in a continuous manner, it gets proportionally added to the part that is to the left of the 12 pm boundary.

Now we are ready to solve the original puzzle. Your degrees of belief in that case can be pictured as being distributed over possible times in two possible worlds: see Fig. 10.2. The development is now such that when the bottom part of the degree of belief distribution hits midnight, it gets snuffed out to the right of midnight, and the rest of the degree of belief distribution is continuously re-normalized, i.e. the top part of the degree of belief distribution and the remaining bottom part are continuously proportionally increased as time passes. Note that Fig. 10.2 is essentially different from Fig. 10.1. In Fig. 10.2 the top distribution starts to increase its absolute size once the leading edge of the bottom distribution hits midnight. This does not happen in Fig. 10.1, since there the degree of belief distributions each were total degree of belief distributions in separate scenarios. Also, in Fig. 10.2 the bottom distribution starts to increase in size once its leading edge hits midnight, but it only increases half as much as it does in Fig. 10.1, since half of the "gains" is being diverted to the top degree of belief distribution.

Thus, at the very least until it actually is midnight, the top and the bottom degree of belief distribution will always be identical to each other, in terms of shape and size, to the left of midnight. Prior to midnight, your degrees of belief will be such that conditional upon it being prior to midnight, it is equally likely to

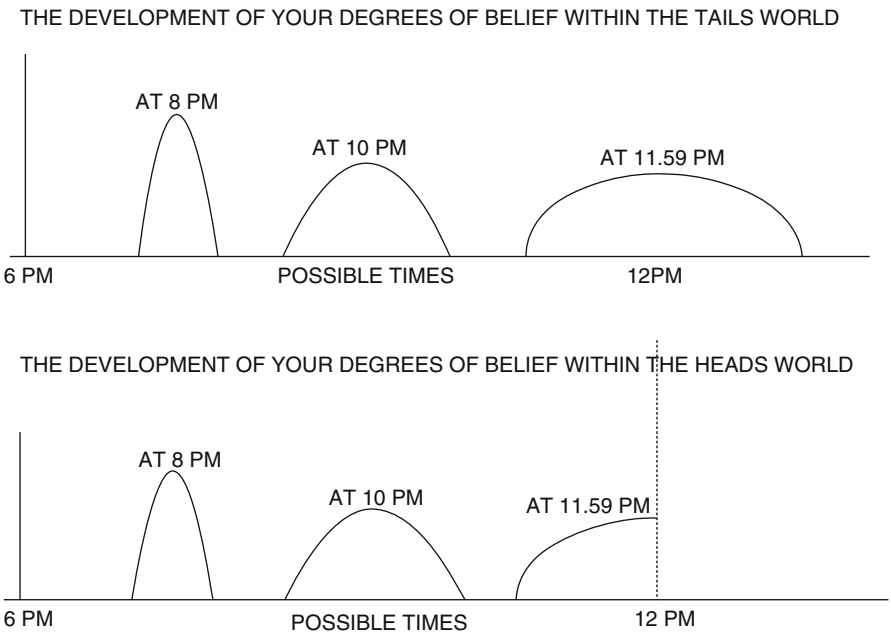


Fig. 10.2 Prisoner with evidence

be heads as tails. Your unconditional degree of belief in tails, however, will increase monotonically as you approach midnight.

After midnight there are two possible ways in which your degree of belief distribution can develop. If the light is switched off your degree of belief distribution collapses completely onto midnight and onto the heads world. If in fact it is not switched off your degree of belief distribution continues to move to the right in both worlds, and it continues to be snuffed out in the heads world to the right of midnight, and the remaining degrees of belief keep being proportionally increased.[1]

Now I can answer the questions that I started with. It is true, as I surmised, that your degree of belief in tails will have increased by 11.59 pm. You will take your internal sense of the passing of time, and combine it with the fact that the light is still on, and you will take this as providing some evidence that the outcome is tails. It is also true, as I surmised, that the light still being on will be taken by you as providing some evidence that it is not yet midnight. For at 11.59 pm your degree of belief distribution over possible times (averaged over the heads and tails worlds) will be further to the left that it would have been had you believed that the light would stay on no matter what. More generally, we have found a unique solution

[1]Thus, for instance, if the light is not switched off, there must be a moment (which could be before or after midnight) such that you have equal degree of belief in each of the 3 possibilities: heads and it is before midnight, tails and it is before midnight, tails and its after midnight.

to the puzzle of how a rational person's sense of time must interact with evidence, given how that person's sense of time works in the absence of evidence.

Rather surprisingly, this interaction can be such, as it is in my example, that you know in advance that at some specified later time you will, almost certainly, have increased your degree of belief in tails, and that you could not possibly have decreased your degree of belief in tails.[2] It is also interesting to note that nothing essential changes in this example if one assumes that the coin toss will take place exactly at midnight. Thus it can be the case that one knows in advance that one will increase one's degrees of belief that a coin toss, which is yet to occur, will land tails. Of course, at the time that one has this increased degree of belief one does not know that this coin toss is yet to occur. Nonetheless, such predictable increases in degrees of belief seem very strange.

John Collins' Prisoner

John Collins has come up with the following variation of the case of the prisoner that was described in the previous section. In Collins's variation the prisoner has 2 clocks in his cell, both of which run perfectly accurately. However, clock A initially reads 6 pm, clock B initially reads 7 pm. The prisoner knows that one of the clocks is set accurately, the other one is one hour off. The prisoner has no idea which one is set accurately; indeed he initially has degree of belief ½ that A is set accurately, and degree of belief ½ that B is set accurately. As in the original case, if the coin lands heads the light in his cell will be turned off at midnight, and it will stay on if it lands tails. So initially the prisoner has degree of belief 1/4 in each of the following 4 possible worlds:

W_1: Heads and clock A is correct
W_2: Heads and clock B is correct
W_3: Tails and clock A is correct
W_4: Tails and clock B is correct.

[2]One might wonder why I inserted the phrase "almost certainly" in this sentence. The reason for this is that there is a subtlety as to whether you know at 6 pm that you will have an increased degree of belief in tails at 11.59 pm. There is an incoherence in assuming that at 6 pm you know with certainty what your degree of belief distribution over possible times will be at 11.59 pm. For if you knew that you could simply wait until your degree of belief distribution is exactly like that. (You can presumably establish by introspection what your degree of belief distribution is.) And when you reach that distribution, you would know that it has to be 11.59 pm. So when that happens you should then collapse your degree of belief distribution completely on it being 11.59 pm. But this is incoherent. Thus, the fact that you do not have a perfect internal clock also implies that you can not know in advance what your degree of belief distribution is going to look like after it has developed (guided only by your internal clock). Thus you can not in advance be certain how your degree of belief distribution over possible times will develop. Nonetheless you can be certain at 6 pm that your degree of belief in tails will not decrease prior to midnight, and that it is extremely likely to have increased by 11.59 pm. At 6 pm your expectation for you degree of belief in tails at 11.59 pm will be substantially greater than 0.5.

When in fact it is 11.30 pm the light, for sure, will still be on. What will the prisoner's degrees of belief then be? Well, if the actual world is W_1, then, when it actually is 11.30 pm clock A will read 11.30 pm and clock B will read 12.30 am. In that case, since the prisoner sees that the light is still on, he will know that it can not be that the coin landed heads and clock B is correct. I.e. his degree of belief in W_2 will be 0, and his degrees of belief in the three remaining options will be 1/3 each. Similarly if the actual world is W_3 then at 11.30 pm prisoner will have degree of belief 0 in W_2 and degree of belief in 1/3 each of the remaining options. On the other hand if the actual world is W_2 or W_4, then when it is actually 11.30 pm, the clock readings will be 10.30 pm and 11.30 pm, and the prisoner will still have the degrees of belief that he started with, namely 1/4 in each of the 4 possibilities. The prisoner, moreover, knows all of this in advance.

This is rather bizarre, to say the least. For, in the first place, at 6 pm the prisoner knows that at 11.30 pm his degrees of belief in heads will be less or equal to what they now are, and can not be greater. So his current expectation of what his degrees of belief in heads will be at 11.30 pm, is less than his current degree of belief in heads. Secondly, there is a clear sense in which he does not trust his future degrees of belief, even though he does not think that he is, or will be, irrational, and even though he can acquire new evidence (the light being on or off). Let D_t denote the prisoner's degrees of belief at time t. Then, e.g., $D_{6.00}$(clock B is correct/$D_{11.30}$(clock B is correct) = 1/3) = 0. For $D_{11.30}$(clock B is correct) = 1/3 only occurs in worlds W_1 and W_3, and in each of those worlds clock B is not correct, and the prisoner knows this. Thus his current degrees of belief conditional upon his future degrees of belief do not equal those future degrees of belief. So he systematically distrusts his future degrees of belief. Strange indeed.

Sleeping Beauty

Some researchers are going to put Sleeping Beauty, SB, to sleep on Sunday night. During the two days that her sleep will last the researchers will wake her up either once, on Monday morning, or twice, on Monday morning and Tuesday morning. They will toss a fair coin Sunday night in order to determine whether she will be woken up once or twice: if it lands heads she will be woken upon Monday only, if it lands tails she will be woken up on Monday and Tuesday. After each waking, she will be asked what her degree of belief is that the outcome of the coin toss is heads. After she has given her answer she will be given a drug that erases her memory of the waking up, indeed it resets her mental state to the state that it was in on Sunday just before she was put to sleep. Then she is put to sleep again. The question now is: when she wakes up what should her degree of belief be that the outcome was heads?

Answer 1: Her degree of belief in heads should be 1/2. It was a fair coin and she learned nothing relevant by waking up.

Answer 2: Her degree of belief in heads should be 1/3. If this experiment is repeated many times, approximately 1/3 of the awakenings will be heads-awakenings, i.e. awakenings that happen on trials in which the coin landed heads.

Adam Elga[3] has argued for the second answer. I agree with him, and I agree with his argument. But let me amplify this view by giving a different argument for the same conclusion. Suppose that SB is a frequent and rational dreamer. Suppose in fact that every morning if SB is not woken up at 9 am, she dreams at 9 am that she is woken up at 9 am. Suppose that the dream and reality indistinguishable in terms of her experience, except that if SB pinched herself and she are dreaming, it does not hurt (and she doesn't wake up), while if she does this while she is awake it does hurt. And let us suppose that SB always remembers to pinch herself a few minutes after she experiences waking up (whether for real, or in a dream.) What should her degrees of belief when she experiences waking up? It seems obvious she should consider the 4 possibilities equally likely (the 4 possibilities being: Monday&Tails&Awake, Monday&Heads&Awake, Tuesday&Tails&Awake, Tuesday&Heads&Dreaming). If SB then pinches herself and finds herself to be awake, she should conditionalize and then have degree of belief 1/3 in each of the remaining 3 possibilities (Monday&Tails&Awake, Monday&Heads&Awake, Tuesday&Tails&Awake). Suppose now that at some point in her life SB loses the habit of dreaming. She no longer needs to pinch herself; directly upon waking she knows that she is not asleep. However, it seems clear that this lack of dreaming should make no difference as to her degrees of belief upon realizing that she is awake. The process now occurs immediately, without the need for a pinch, but the end result ought to be the same.

Here again the crucial assumption is commutativity: if the relevant evidence and experience collected is the same, then the order of collection should not matter for the final degrees of belief.[4] But there is clearly something very puzzling about such foreseeable changes in degrees of belief.

Duplication

Scenario 1 While you are at the beach, Vishnu tells you that, contrary to appearances, you have existed only for one month: Brahma created you one month ago, complete with all your memories, habits, bad back, and everything. What's more, says Vishnu, one month ago Brahma in fact created two human beings like you (you are one of them), in exactly the same environment, at two different ends of

[3]Elga, A. (2000): "Self-locating belief and the Sleeping Beauty problem", *Analysis* **60**, pp 143–147.

[4]Cian Dorr has independently arrived at the idea of using commutativity in order to argue for the degrees of belief that Elga advocates in the Sleeping Beauty case. See Dorr, C.: "Sleeping Beauty: in defence of Elga", forthcoming, *Analysis*.

the universe: one on earth, one on twin earth. Unfortunately, Vishnu has a further surprise for you: one month ago Shiva tossed a coin. If it landed heads Shiva will destroy the human being that is on twin earth one month from now. If it landed tails Shiva will do nothing. Vishnu does no tell you whether you are to be destroyed, but recommends that if you want to know, you should go check your mail at home. If there is a letter from president Bush for you, then you will be destroyed. Before running home, what degree of belief should you have in the 4 possibilities: Earth&Heads, Earth&Tails, Twin Earth&Heads, Twin Earth&Tails? It seems clear that you should have degree of belief 1/4 in each, or at the very least, that it is not irrational to have degree of belief 1/4 in each. You run home, and find no letter from Bush. What should your degrees of belief now be? Well, by conditionalization, they should now be 1/3 in each of the remaining possibilities (Earth&Tails, Twin Earth&Heads, Twin Earth&Tails). Consequently you should now have degree of belief 1/3 that the toss landed heads and 2/3 that it landed tails.

Scenario 2 same as scenario 1, except that Vishnu tells you that if the toss came heads, your identical twin was destroyed by Shiva one week ago. Since you were obviously not destroyed, you do not need to rush home to look for a letter from Bush. In essence you have learned the same as you learned in the previous scenario when you found you had no letter from Bush, and hence you should now have degree of belief 1/3 that the toss landed heads.

Scenario 3 same as scenario 2, except that Vishnu tells you that rather than that 2 beings were created one month ago by Brahma, one of them already existed and had the exactly the life you remember having had. This makes no relevant difference and you should now have degree of belief 1/3 that the coin landed heads.

Scenario 4 same as scenario 3, except that Vishnu tells you that if the die landed heads one month ago Shiva immediately prevented Brahma from creating the additional human being one month ago. The upshot is that only if the coin landed tails Brahma will have created the additional human being. Since the timing of the destruction/prevention makes no relevant difference you should again have degree of belief 1/3 that the coin landed heads.

Scenario 5[5] You are on earth, and you know it. Vishnu tells you that one month from now Brahma will toss a coin. If it lands tails Brahma will create, at the other end of the universe, another human being identical to you, in the same state as you will then be, and in an identical environment as you will then be. What do you now think that your degrees of belief should be in one month time? The answer is that they should be the same as they are in scenario 5, since in one month time you will be in exactly the epistemic situation that is described in scenario 5. Of course, it is plausible to claim that your future self will actually be on earth, since it is only your future continuation on earth that can plausibly be called "your future self". However, that does not mean that your future self can be sure that he is on earth. For your future self will know that he will have the same experiences and memories,

[5]This scenario is similar to the "Dr Evil scenario" in Elga, A. (manuscript): "Defeating Dr. Evil with self-locating belief".

whether or not he is on earth or on twin earth, and thus he will not know whether he can trust his memories. Thus you now have degree of belief ½ in heads, and yet you know that in one month time, you will have degree of belief 1/3. This is bizarre, to say the least.

Yet again, the crucial assumption in this reasoning is commutativity: your final degrees of belief should not depend on the order in which you receive all the relevant experience and evidence. You should end up with the same degrees of belief, namely degree of belief ½ in heads, whether you all along knew you were on Earth, or whether you only later found out that you were on Earth. But that can only be so if you had degree of belief 1/3 in heads prior to discovering that you were on Earth.

Diagnosis

Bas van Fraassen's Reflection Principle[6] says that one should trust one's future degrees belief in the sense that one's current degree of belief D_0 in any proposition X, given that one's future degree of belief D_t in X equals p, should be p: $D_0(X/D_t(X) = p) = p$. Given that one is sure that one will have precise degrees of belief at time t, the Reflection Principle entails that one's current degrees of belief equal the expectations of one's future degrees of belief: $D_0(X) = \sum pD_0(D_t(X) = p)$. The Reflection Principle is violated in each of the 5 puzzles that I have presented, for in each case there is a time at which one's expectation of one's future degree of belief in Heads differs from one's current degree of belief in Heads. This is presumably why we find these cases, prima facie, so worrying and strange.

The source of the problem, I claim, is that the degrees of belief of perfectly rational people, people who are not subject to memory loss or any other cognitive defect, can develop in ways that are as yet unrecognized, and indeed are not allowed according to standard Bayesian lore. Standard Bayesian lore has it that rational people satisfy the Principle of Conditionalization: rational people alter their degrees of belief only by strict conditionalization on the evidence that they acquire.[7] Strict conditionalization of one's degrees of belief upon proposition X can be pictured in the following manner. One's degrees of belief are a function on the set of possibilities that one entertains. Since this function satisfies the axioms of probability theory it is normalized: it integrates (over all possibilities) to 1.

[6] See van Fraassen (1995): "Belief and the problem of Ulysses and the sirens", *Philosophical Studies* 77: 7–37.

[7] Strict conditionalization: when one learns proposition X at t, one's new degrees of belief D_t equals one's old degrees of belief D_0 conditional upon X: $D_t(Y) = D_0(Y/X)$. One might also allow Jeffrey conditionalization. It matters not for our purposes.

Conditionalizing such a function on proposition X then amounts to the following: the function is set to 0 over those possibilities that are inconsistent with X, while the remaining non-zero part of the function is boosted (by the same factor) everywhere so that it integrates to 1 once again. Thus, without being too rigorous about it, it is clear that conditionalization can only serve to 'narrow down' one's degree of belief distribution (one really learns by conditionalization). In particular a degree of belief distribution that becomes more 'spread out' as time passes can not be developing by conditionalization, and a degree of belief distribution that exactly retains its shape, but is shifted as a whole over the space of possibilities, can not be developing by conditionalization. However, such spreading out and shifting is exactly what occurs in the 5 puzzles that I presented.

The reasons for such spreading and shifting are very simple. First let us consider shifting. Suppose that one knows exactly what the history of the world that one inhabits is like. And suppose that one is constantly looking at a clock one knows to be perfect. One's degrees of belief will then be entirely concentrated on one possible world, and at any given moment one's degrees of belief within that world will be entirely concentrated on one temporal location, namely the one that correspond to the clock reading that one is then seeing. And that of course means that the location that one's degree of belief distribution is concentrated at is constantly moving. That is to say, one's degree of belief distribution is constantly shifting, and such a constant shifting is simply not a case of Conditionalization. Self-locating beliefs will therefore genetically develop in ways that violate Conditionalization. John Collins's prisoner case involves exactly such a shifting of one's self-locating degrees of belief. The only difference is that in his case one additionally has an initial uncertainty as to which clock is accurate, i.e. one is initially uncertain whether one is in a world in which clock A is correct or one in which clock B is correct. It is somewhat surprising that this kind of violation of Conditionalization can be parlayed into a violation of Reflection. But Collins' prisoner case shows exactly how one can do this.

Next let us consider spreading. The simplest case of spreading is the case of the traveller that takes the path by the Mountains to Shangri La. His degrees of belief become more spread out when he arrives in Shangri La: at that time he goes from degrees of belief 1 in Heads and 0 in Tails, to degrees of belief ½ in Heads and ½ in Tails.[8] The reason why this happens is that there are two distinct possible

[8]Bas van Fraassen has, in conversation with me, has suggested that in such situations Conditionalization indeed should be violated, but Reflection should not. In particular he suggested that the degrees of belief of the traveler should become completely vague, upon arrival in Shangri La. This does not strike me as plausible. Surely upon arrival in Shangri La our traveler is effectively in the same epistemic situation as someone who simply knows that a fair coin has been tossed. One can make this vivid by considering two travelers, A and B. Traveler A never looks out of the window of the car, and hence maintains degree of belief ½ in heads all the way. (The memory replacement device does not operate on travelers who never look out of the window.) Traveler A, even by van Fraassen's lights, upon arrival in Shangri La, should still have degree of belief ½ in Heads. However, traveler B, does look out of the window during the trip. Upon arrival, by van Fraassen's lights, B's degrees of belief should become completely vague. But it seems odd to me that traveler B is epistemically penalized, i.e. is forced to acquire completely vague degrees of

experiential paths that end up in the same experiential state. That is to say the traveler's experiences earlier on determine whether possibility A is the case (Path by the Mountain), or whether possibility B is the case (Path by the Ocean). But because of the memory replacement that occurs if possibility B is the case, those different experiential paths merge into the same experience, so that that experience is not sufficient to tell which path was taken. Our traveler therefore has an unfortunate loss of information, due to the loss of the discriminating power of his experience. What is somewhat surprising is that this loss of discriminating power is not due to any loss of memory or any cognitive defect on his part: it is due to the fact that something strange would have happened to him had he taken the other path! This loss of discriminatory power of experience, and consequent spreading out of degrees of belief here does not involve self-locating degrees of belief. Suppose, e.g., that our traveler is the only person ever to travel along either paths. Then our traveler initially is unsure whether he is in a world in which path A is never taken or whether he is in a world in which path B is never taken. He then becomes sure that he is in a world in which path B is never taken. Even later, upon arrival, he again becomes unsure as to which world he is in. None of this has anything to do with self-locating beliefs.[9]

The source of the Sleeping Beauty and Duplication problems is exactly the same. In the case of Sleeping Beauty the possibility of memory erasure ensures that the self-locating degrees of belief of Sleeping Beauty, even on Monday when she has suffered no memory erasure, become spread out over two days. In the Duplication case, yet again, the possible duplication of experiences forces one to become uncertain as to where (or who) one is. The cause of the spreading of degrees of belief in both cases is "experience duplication", and has nothing to do with the self-locating nature of these beliefs.[10]

It is not very surprising that the spreading of degrees of belief can bring about a violation of Reflection. For instance, in the non-self locating case a predictable reduction from degree of belief 1 in some proposition X to anything less than 1 will

belief, just because he looked out of the window during the trip, when it seems clear that he ends up in exactly the same epistemic position as his companion, who did not look out of the window.

[9]It is obvious how to generalize this case to a case in which there are memory replacement devices at the end of both roads, where these memory replacement devices are indeterministic, i.e. when it is the case that for each possible path there are certain objective chances for certain memories upon arrival in Shangri La. For, given such chances (and the Principal Principle), one can easily calculate the degrees of belief that one should have (in heads and tails) given the memory state that one ends up with. And, generically, one will still violate Conditionalization and Reflection.

[10]Some people will balk at some of the degrees of belief that I have argued for in this paper, in particular in the self-locating cases. For instance, some people will insist that tomorrow one should still be certain that one is on Earth, even when one now knows (for sure) that a perfect duplicate of oneself will be created on Mars at midnight tonight. I beg to differ. However, even if in this case, and other cases, one disagrees with me as to which degrees of belief are rationally mandated, the main claim of this paper still stands. The main claim is that in such cases of possible experience duplication, it is at the very least rationally permissible that one's degrees of belief become more spread out as time progresses, and hence rational people can violate Conditionalization and Reflection.

immediately violate Reflection: now you know it, now you don't. The argument is slightly less straightforward in the self-locating case. Consider, e.g., a case in which one is on Earth and one knows that at midnight a duplicate of oneself will be created on Mars. One might claim that since one now is certain that one is on Earth, and at midnight one will be uncertain as to whether one is on Earth, that one has a clear violation of Reflection. However, this is too quick. To have a clear violation of Reflection it has to be the very same "content of belief" such that one's current degree of belief differs from one's expectation of one's future degree of belief. Depending on what one takes to be the contents of belief when it concerns self-locating beliefs (propositions? maps from locations to propositions? .....?), one might argue that the contents of belief are not the same at the two different times, and hence there is no violation of Reflection. However the arguments of sections IV and V show that one can in any case parlay such spreading of self-locating degrees of belief into violations of Reflection concerning such ordinary beliefs as to whether a coin lands Heads or Tails. So Reflection is suckered anyhow.

Finally, the original case of the prisoner involves both a spreading of degrees of belief and a shifting of degrees of belief. The shifting is due simply to the passage of time and the self-locating nature of the beliefs. The spreading is due to the fact that our prisoner does not have experiences that are discriminating enough to pick out a unique location in time.[11] The analysis of section II shows, yet again, that such a spreading and shifting of self-locating degrees of belief can be parlayed into a violation of Reflection concerning such ordinary beliefs as to whether a coin lands Heads or Tails.

Conclusions

The degrees of belief of rational people can undergo two as yet unrecognized types of development. Such degrees of belief can become more spread out due to the duplication of experiences, or more generally, due to the loss of discriminating power of experiences, and thereby violate Conditionalization. In addition self-locating degrees of belief will generically be shifted over the space of possible

[11] One might model the prisoner here as having unique distinct experiences at each distinct, external clock, time, and as initially having precise degrees of belief over the possible ways in which those experiences could correlate to the actual, external clock, time. If one were to do so then the prisoner would merely be initially uncertain as to which world he was in (where worlds are distinguished by how his experiences line up with the actual, external clock, time), but for each such possible world would be always certain as to where he was located in it. And, if one were to do so, then the original prisoner case would be essentially the same case as Collins's prisoner case: no uncertainty of location in any given world, merely an initial uncertainty as to which world one is in, and a subsequent shifting of the locally concentrated degrees of belief within each of the possible worlds. However, there is no need to represent the original prisoner case that way. Indeed it seems psychologically somewhat implausible to do so. More importantly, the arguments and conclusions of this paper do not depend on how one models this case.

locations, due to the passage of time, and thereby violate Conditionalization. Such violations of Conditionalization can be parlayed into violations of Reflection, and lead to a distrust of one's future degrees of belief. Strange, but not irrational.

Acknowledgements I would like to thank John Collins, Adam Elga, John Hawthorne, Isaac Levi, Barry Loewer, and Tim Maudlin for extensive and crucial comments and discussions on earlier versions of this paper.

Chapter 11
Stopping to Reflect

Mark J. Schervish, Teddy Seidenfeld, and Joseph B. Kadane

Our note is prompted by a recent article by Frank Arntzenius, "Some Problems for Conditionalization and Reflection".[1] Through a sequence of examples, that article purports to show limitations for a combination of two inductive principles that relate current and future rational degrees of belief: *Temporal Conditionalization* and *Reflection*:

(i) *Temporal Conditionalization* is the rule that, when a rational agent's corpus of knowledge changes between *now* and *later* solely by learning the (new) evidence, B, then coherent degrees of belief are updated using conditional probability according the formula, for each event A,

$$P_{later}(A) = P_{later}(A|B) = P_{now}(A|B)$$

(ii) *Reflection*[2] between *now* and *later* is the rule that current conditional degrees of belief defer to future ones according to the formula that, for each event A,[3]

$$P_{now}(A|P_{later}(A) = r) = r.$$

[1] *The Journal of Philosophy* Vol C, Number 7 (2003), 356–370.

[2] See B.van Frasseen's "Belief and the Will," this *Journal*, 81 (1984), 235–256. van Fraassen's *Reflection* has an antecedent in M.Goldstein's "Prevision of a Prevision," *JASA* 78 (1983): 817–819.

[3] Here and through the rest of this note 'r' is a standard designator for a real number – this in order to avoid Miller-styled problems. See, D.Miller's "A Paradox of Information," *Brit. J. Phil. Sci*. 17 (1966):144–147.

M.J. Schervish (✉) • T. Seidenfeld • J.B. Kadane
Carnegie Mellon University, Pittsburgh, PA, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_11

We will use the expression "*Reflection* holds with respect to the event A." to apply to this equality for a specific event A.

It is our view that neither of these principles is mandatory for a rational agent.[4] However, we do not agree with Arntzenius that, in the examples in his article, either of these two is subject to new restrictions or limitations beyond what is already assumed as familiar in problems of stochastic prediction.

To the extent that a rational person does not know *now* exactly what she or he will know in the future, anticipating one's future beliefs involves predicting the outcome of a stochastic process. The literature on stochastic prediction relies on two additional assumptions regarding states of information and the temporal variables that index them[5]:

(iii) When $t_2 > t_1$ are two fixed times, then the information the agent has at t_2 includes all the information that she or he had at time t_1.[6] This is expressed mathematically by requiring that the collection of information sets at all times through the future form what is called a *filtration*.

Second, since the agent may not know *now* the precise time at which some specific information may become known in the future, then future times are treated as *stopping times*. That is:

(iv) For each time T (random or otherwise) when a prediction is to be made, the truth or falsity of the event $\{T \leq t\}$ is known at time t, for all fixed t. Such (random) times T are called *stopping times*.

In this note, we apply the following three results[7] to the examples in Arntzenius' article. These results, we believe, help to explain why the examples at first appear puzzling and why they do not challenge either *Temporal Conditionalization* or *Reflection*. Result 11.1 covers the ordinary case, where *Reflection* holds. Results 11.2 and 11.3 establish that *Reflection* will fail when one or the other of the

[4] We have argued, for example, that when (subjective) probability is finitely but not countably additive, then there are simple problems where (*i*) is reasonable, but where (*i*) precludes (*ii*). See our "Reasoning to a Foregone Conclusion," *JASA* 91 (1996): 1228–1236. Also, Levi argues successfully, we think, that (*i*) is not an unconditional requirement for a rational agent. See his "The Demons of Decision," *The Monist* 70 (1987): 193–211.

[5] See, for example, section 35 of P.Billingsley, *Probability and Measure* 3*rd* edition, J.Wiley, 1995.

[6] Here and through the rest of this note 't' is a standard designator for a real number for time. More precisely, we use the subscripted variable, e.g. 't_1' to denote a specific time as the agent of the problem is able to measure it. We presume that the agent has some real-valued "clock" that quantifies a transitive relation of "is later than." Subtleties about the differences between how time is so indexed for different observers is relevant to one of Arntzenius' puzzles, to wit, the Prisoner's Problem.

[7] Proofs for these three are given in the "Appendix". In this note, we assume that all probability is countably additive.

two additional assumptions, (*iii*) and (*iv*) fail. It is not hard to locate where these assumptions are violated in the examples that Arntzenius presents.

Result 11.1 When "*later*" is a *stopping time*, when the information sets of future times form a *filtration*, and assuming that degrees of belief are updated by *Temporal Conditionalization*, then *Reflection* between *now* and *later* follows.

Result 11.2 When the information known to the agent over time *fails* to form a filtration, not only is *Temporal Conditionalization* vacuously satisfied (as its antecedent fails), but then *Reflection* fails unless what is forgotten in the failure of filtration becomes practically certain (its probability becomes 0 or 1) in time for future predictions, *later*.

Result 11.3 However, if the information known to the agent over time forms a filtration and *Temporal Conditionalization* holds, but "*later*" is not a *stopping time*, then *Reflection* between *now* and *later* holds for the specific event A, i.e., $P_{now}\,(A|P_{later}(A) = r) = r$, subject to the necessary and sufficient condition, (11.1), below.

Let H_t be the event "$t = later$." When *later* is not a stopping time, the event H_t is news to the agent making the forecasts. The question at hand is whether this news is relevant to the forecasts expressed by *Reflection*. To answer that question, concerning such forecasts about the event A, define the quantity $y_t(A)$ by

$$y_t(A) = \frac{P_{now}\,(H_t|P_t(A) = r\&A)}{P_{now}\,(H_t|P_t(A) = r)}.$$

The quantity $y_t(A)$ is an index of the current conditional dependence between A and H_t, given that $P_t(A) = r$. For example, $y_t(A) = 1$ if and only if A and H_t are conditionally independent for the agent, *now*, given that $P_t(A) = r$. In other words, by symmetry of conditional independence, $y_t(A) = 1$ if and only if the agent's current conditional probability of A given that $P_{later}(A) = r$, is unchanged by the added information H_t.

Reflection holds for A between *now* and *later*, $P_{now}\,(A|P_{later}(A) = r) = r$ if and only if, given $P_{later}(A) = r$, the conditional expected value $y_T(A) = 1$. Specifically, if and only if

$$1 = \sum\nolimits_t y_t(A) P_{now}\,(H_t|P_{later}(A) = r) \tag{11.1}$$

Thus, *Reflection* is satisfied between *now* and *later* if and only if (11.1) holds for each A.

Next, we illustrate the second and third results with examples that show how *Reflection* may fail.

Example 11.1 (Illustrating Result 11.2)

Suppose that the agent will observe a sequence of coin tosses, one at a time at a known rate, e.g. one toss per minute. Let $X_n = 1$ if the coin lands heads up on toss n, and let $X_n = 0$ otherwise. The agent does not know how the coin is loaded, but believes that it is *fair* (event A) with personal probability ½, and that with personal probability ½ it is biased with a chance of ¾ for landing tails (event A^c). Also, he believes that tosses are conditionally independent given the loading, i.e., given that the coin is *fair* or given that it is biased ¾ for tails.

Time is indexed for the agent by the number of the most recent coin toss. The time "*now*" occurs after the first toss ($t = n = 1$), and "*later*" denotes the time ($t = n = 2$) just after the second toss. Unfortunately, at each time t, the agent knows that he can remember only the most recent flip, X_t, though he knows which numbered toss it is because, e.g., he can see a clock. Suppose that the first toss lands heads up, which is the event $C = \{X_1 = 1\}$. The information that will be available to the forgetful agent later (at $t = 2$) will be only that either $B_1 = \{X_2 = 1\}$ or $B_0 = \{X_2 = 0\}$. He will not recall C because of his predictable memory lapse, and he knows all this. It is straightforward to compute:

$$P_{later}(A|B_1) = 2/3 \text{ and } P_{later}(A|B_0) = 2/5.$$

However, at $t = 1$, the agent's conditional probability for A, given event B_1 occurring at $t = 2$, satisfies $P_{now}(A \mid B_1) = 4/5$. Similarly, if *now* he conditions on event B_0 occurring at $t = 2$, his conditional probability will satisfy $P_{now}(A \mid B_0) = 4/7$.

Of course, *Temporal Conditionalization* holds vacuously at the *later* time, since the information sets available to the agent do not form a filtration. *Reflection* fails in this setting, as the agent does not remember at the *later* time what happened *now*, and he knows this all along. If B_1 occurs then $P_{later}(A) = P_{later}(A|B_1) = 2/3$, and if B_0 occurs then $P_{later}(A) = P_{later}(A|B_0) = 2/5$. Hence,

$$P_{now}(A|P_{later}(A) = 2/3) = 4/5$$

and

$$P_{now}(A|P_{later}(A) = 2/5) = 4/7. \qquad \square$$

Example 11.2 (Illustrating Result 11.3 when condition (11.1) fails and then *Reflection* fails too)

Modify Example 11.1 so that the agent has no memory failures and updates his degrees of belief by *Temporal Conditionalization*. Also, change the time "*now*" to denote the minute prior to the first toss, i.e., *now* is $t = n = 0$. Define the time "*later*" to be the *random time, T, just prior to the first toss that lands heads up*. From the point of view of the agent, the quantity T is not an observable random variable up to and including time T, and it is not a *stopping time* either. It is observable to the agent starting with time $T + 1$, of course, as by then he will have seen when the first head occurs.

With probability 1 the possible values for T are $T = 0, 1, 2, \ldots$. It is straightforward to verify that: $P_{later}(A) = [1 + (3/2)^n]^{-1}$, when $T = n$, for $n = 0, 1, 2,$

. . . . Notice that $P_{later}(A) \leq ½$, no matter when T occurs, and $P_{later}(A) < ½$ for $T > 0$, since if $T > 0$, the initial sequence of tosses that the agent observes all land tails up. However, from the value of $P_{later}(A)$ and knowing it is this quantity, one may calculate T exactly and thus know the outcome of the $n + 1$st toss, which is a heads. But when the agent computes $P_{later}(A)$ at the time *later*, he does not then know that *later* has arrived. Thus, *later*, he is not in a position to use the extra information that he would get from knowing when T occurs to learn the outcome of the $n + 1$st toss. To repeat the central point, T is not a stopping variable.

It is evident that *Reflection* fails, $P_{now}(A \mid P_{later}(A) = r) \neq P_{later}(A)$. The extra information, namely that $P_{later}(A) = r$ rather than merely that $P_t(A) = r$ where t is the time on the agent's clock, is information that is relevant to his *current* probability of A, since it reveals the outcome of the next toss. Even *now*, prior to any coin tosses, when he computes $P_{now}(A \mid P_{later}(A) = r)$, the conditioning event reveals to him the value of T, since n is a function of r. In this case, the conditioning event entails the information of n and when the first heads occurs, namely, on the $n + 1$st toss. Then *Reflection* fails as

$$P_{now}\left(A|P_{later}(A) = [1 + (3/2)^n]^{-1}\right) = \left(1 + 3^n/2^{n+1}\right)^{-1}.$$

It remains only to see that (11.1) fails as well. Consider the quantity $y_t(A)$ used in condition (11.1). $y_t(A) = \frac{P_{now}(H_t|P_t(A)=r\&A)}{P_{now}(H_t|P_t(A)=r)}$. Given $P_t(A) = r$, the added information that A obtains is relevant to the agent's current probability when *later* occurs. Specifically, as $P_t(A) = [1 + (3/2)^n]^{-1}$ entails that $t = n$,

$$\begin{aligned} P_{now}\left(H_t|P_t(A) = [1 + (3/2)^n]^{-1}\right) &= P_{now}\left(X_{t+1} = 1|P_t(A) = [1 + (3/2)^n]^{-1}\right) \\ &= (1/2)\,[1 + (3/2)^n]^{-1} + (1/4)\,(3/2)^n[1 + (3/2)^n]^{-1} < \tfrac{1}{2} \\ &= P_{now}\left(X_{t+1} = 1|P_t(A) = [1 + (3/2)^n]^{-1}\&A\right) \\ &= P_{now}\left(H_t|P_t(A) = [1 + (3/2)^n]^{-1}\&A\right). \end{aligned}$$

Thus, $y_t > 1$.

Hence, $1 < \sum_t y_t(A) P_{now}\,(H_t \mid P_{later}(A) = r)$. □

Example 11.3 (Illustrating Result 11.3 when (11.1) obtains and *Reflection* holds even though *later* is not a *stopping time*)

In this example, consider a sequence of three times, $t = 0$, 1, and 2. *Now* is time $t = 0$. The available information increases with time, so that the information sets form a filtration, and the agent updates his degrees of belief by *Temporal Conditionalization*. Let the random time *later* be one of the two times $t = 1$, or $t = 2$, chosen at random, but which one is not revealed to the agent. Let the event H_i be that $later = i$, $(i = 1, 2)$ and suppose that the occurrence of H_i (or its failure) while not known to the agent at any of the three times is independent of all else that the agent does know at all three times. In this case, for each event A (even for $A = H_i$) Eq. (11.1) is satisfied. That is, by the assumptions of the problem, either

$y_t(A) = \frac{P_{now}(H_t|y_t(A)=r\&A)}{P_{now}(H_t|P_t(A)=r)} = 1$, or if $A = H_i$ then $y_t(A) = \frac{P_{now}(H_t|P_{later}(A)=r)}{P_{now}(H_t|P_t(A)=r)} = 1$. Thus, $P_{now}(A \mid P_{later}(A) = r) = r$. That is, even though *later* is not a stopping time, *Reflection* holds in this case since, given that $P_{later}(A) = r$ no new (relevant) evidence about A is conveyed through knowing that *later* has arrived, H_i. □

We note that Result 11.2 applies to the *Sleeping Beauty*[8] *Shangri La*, and *Duplication* examples of Arntzenius' article, where known failures of memory are explicit to the puzzles. Result 11.3 applies to explain the failure of *Reflection* in the two versions of the "Prisoner" example where the *local time* in the story, as measured by an ordinary clock (e.g., "11:30 PM" in John Collins's example) is not a *stopping time* for the Prisoner.

It is our impression of Collins's *Prisoner* example that the reader is easily mistaken into thinking that the *local time*, as measured by an ordinary clock in the story, is a *stopping time* for all the characters in the story. Then *Reflection* holds for each of them, in accord with Result 11.1. In Collins' example, the *local time*, e.g., 11:30 PM, is a *stopping time* for the Jailor (and also for the reader), but *not* for the Prisoner. For the Prisoner, time is measured by real-valued increments over the starting point, denoted by "now." Increments of *local time* are *stopping times* for the Prisoner. This is because the Prisoner does not know at the start of the story which of two *local times* equals his time *now*. For subsequent times, he does know how much *local time* has elapsed since *now*. But that information is not equivalent to knowing the *local time*. That difference in what is a *stopping time* for different characters is what makes this a clever puzzle, we think.

Acknowledgements Our research was carried out under NSF Grant DMS 0139911. We thank Joseph Halpern for alerting one of us (*T.S.*) to the *Sleeping Beauty* problem, independent of Arntzenius' article.

[8]See also J.Y.Halpern's "Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems," Dept. of Computer Science, Cornell University. September, 2003. We agree with Halpern that, in our words, coherence of a sequence of previsions does not require that they will be *well calibrated* – in a frequency sense of "well calibrated." That is, we think it is reasonable for Sleeping Beauty to give a prevision of ½ to the event that the known fair coin landed heads on the flip in question, each time she is woken up. What complicates the analysis is that the repeated trials in Sleeping Beauty's game do not form an independent sequence, and her mandated forgetfulness precludes any "feedback" about the outcome of past previsions. When repeated trials are dependent and there is no learning about past previsions, coherent previsions may be very badly calibrated in the frequency sense. For other examples and related discussion of this point see, e.g., Seidenfeld, T. (1985) "Calibration, Coherence, and Scoring Rules," *Philosophy of Science* 52: 274–294.

Appendix

Proof of Result 11.1[9] Assume that when X is a random variable and C is an event, the agent's expected value $\mathrm{E_P}(X)$ and conditional expected value $\mathrm{E_P}(X|\ C)$ exist with respect to the probability P. Let A be an event and let $X = P(A|Y)$ be a random variable, a function of the random variable Y. Then, as a consequence of the law of total probability, with C also a function of Y,

$$P\,(A|C) = E_P\,[X|C]\,. \tag{11.2}$$

Assume that the agent's degrees of belief *now* include his *later* degrees of belief as objects of uncertainty. That is, future events such as "$P_{later}(A) = r$" and "$P_{later}(A \mid C) = q$" are proper subjects, *now*, of the agent's current degrees of belief. Suppose that, *now*, the agent anticipates using (*i*) *Temporal Conditionalization* in responding to the new evidence $Y = y$ that he knows he will learn at the stopping time, *later*. For example, Y might be the result of a meter reading made at the *later* time, with a sample space of m many possible values $Y = \{y_1,\ \ldots,\ y_m\}$. Thus, by (*i*), for whichever value y of Y that results,

$$P_{later}(A) = P_{later}\,(A|Y = y) = P_{now}\,(A|Y = y)\,. \tag{11.3}$$

Then, by (*i*) and (11.2), for C also a function of Y, the agent *now* believes that

$$P_{now}\,(A|C) = E_{P_{now}}\,[P_{later}(A)|C]\,. \tag{11.4}$$

Let C be the event, "$P_{later}(A) = r$," which we presume is a possible value for $\mathrm{P}_{later}(A)$ from the agent's current point of view. (This C is function of Y.) Then, because *later* is a stopping time,

$$P_{now}\,(A|P_{later}(A) = r) = E_{P_{now}}\,[P_{later}(A)|P_{later}(A) = r)]. \tag{11.5}$$

As

$$\mathrm{E}_{P_{now}}\,[P_{later}(A)|P_{later}(A) = r)] = r, \tag{11.6}$$

[9] van Fraassen (1995) "Belief and the Problem of Ulysses and the Sirens," *Phil. Studies* 77: 7–37, argues (pp. 17–19) that *Temporal Conditionalization* implies *Reflection*. His argument (pp. 18–19) has an additional, tacit assumption that the time t at which conditioning applies for *Reflection* is a stopping time.

therefore

$$P_{now}\,(\mathrm{A}|P_{later}(A) = r) = r, \tag{11.7}$$

i.e., then *Reflection* holds as well. □

Proof of Result 11.2 To show that *Reflection* fails, consider two times $t_1 < t_2$. Call an event *forgotten* if its truth or falsity is known at time t_1 but not at time t_2. From the assumption that these times do not form a filtration, let E be forgotten between t_1 and t_2 and allow that at t_1 this is known to happen at t_2. Since $P_{t_1}(E) \in \{0, 1\}$, conditioning will not change this value, i.e.,

$$P_{t_1}(E) = P_{t_1}\,(E|P_{t_2}(E) = r) \tag{11.8}$$

for a set of r-values of probability 1 under P_{t_1}. But, since it is known at t_1 that E will be forgotten at t_2, $P_{t_1}\,(0 < P_{t_2}(E) < 1) = 1$. Hence *Reflection* fails as $0 < r < 1$ in (11.8). □

Proof of Result 11.3 Assume that the agent's information sets form a filtration over time and that *Temporal Conditionalization* holds between *now* and *later* but that *later* is not a stopping time for the agent. Let H_t be the event "*later* $= t$" for the specific time t. That is, assume that $0 < P_{later}(H_t) < 1$, when *later* occurs at t.

Later is the future time we will focus on in calculating whether *Reflection* holds, i.e. we will inquire whether for each event A, $P_{now}(\mathrm{A} \mid P_{later}(A) = r) = r$, or not. We calculate as follows.

$$P_{now}\,(A|P_{later}(A) = r)$$

$$= \sum\nolimits_t P_{now}\,(A\&H_t|P_{later}(A) = r)$$

by the law of total probability.

$$= \sum\nolimits_t P_{now}\,(A|P_{later}(A) = r\&H_t)\,P_{now}\,(H_t|P_{later}(A) = r)$$

by the multiplication theorem

$$= \sum\nolimits_t \frac{P_{now}\,(H_t|P_t(A) = r\&A)}{P_{now}\,(H_t|P_t(A) = r)} P_{now}\,(A|P_t(A) = r)\,P_{now}\,(H_t|P_{later}(A) = r)$$

by Bayes' theorem and the equivalence of

$$(P_{later}(A) = r\&H_t)\ \textit{and}\ (P_t(A) = r\&H_t)$$

$$= r\sum\nolimits_t \frac{P_{now}\,(H_t|P_t(A) = r\&A)}{P_{now}\,(H_t|P_t(A) = r)} P_{now}\,(H_t|P_{later}(A) = r)$$

as $P_{now}\,(A|P_t(A) = r) = \mathrm{r}$ by Result 11.1.

$$= r\sum\nolimits_t y_t(A) P_{now}\,(H_t|P_{later}(A) = r).$$

by the definition of $y_t(A)$

Hence, $P_{now}\,(A|P_{later}(A) = r) = r$ if and only if $\sum_t y_t(A) P_{now}(H_t|P_{later}(A) = r) = 1$, which is condition (11.1). □

Part II
Belief Change

Chapter 12
Introduction

Horacio Arló-Costa, Vincent F. Hendricks, and Johan van Benthem

It is well known that the usual versions of probability kinematics have serious limitations. According to the classical notion of conditioning when one learns a piece of information *A* its probability raises to its maximum (one). Moreover no further instance of learning will be capable of defeating *A*. Once a piece of information is learned one should be maximally confident about it and this confidence should remain unaltered forever. It is clear that there are many instances of learning that cannot be accommodated in this Procrustean bed. There are various ways of amending this limited picture by enriching the Bayesian machinery. For example, one can appeal to a notion of primitive conditional probability capable of making sense of conditioning on zero measure events. But the detailed consideration of this alternative leads to similar limitations: the picture of learning that thus arises continues to be cumulative. There are many ways of overcoming these important limitations. Williamson considers one possible way of doing so in his essay reprinted in the section on Bayesian epistemology. One of the lessons that have been learned in recent years is that there is no apparent way of circumventing this rigidity of Bayesianism without introducing in some way a qualitative doxastic or epistemic notion as a primitive alongside probability. Here are two examples:

Horacio Arló-Costa was deceased at the time of publication.

H. Arló-Costa (deceased)
Carnegie Mellon University, Pittsburgh, PA, USA

V.F. Hendricks (✉)
Center for Information and Bubble Studies, University of Copenhagen, Copenhagen, Denmark
e-mail: vincent@hum.ku.dk

J. van Benthem
University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Stanford University, Stanford, United States
e-mail: johan@science.uva.nl

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_12

Williamson proposes a model where knowledge is a primitive, while Levi appeals to a primitive notion of *full belief*.

The traditional work in belief revision has followed Levi in adopting some doxastic notion as a primitive (which does not need to be full belief, it could be Spohn's *plain belief*, etc). But if such a qualitative notion should be introduced primitively how should one go about representing it? And how one would go about characterizing changes of the represented notion? There are many possible options from sentential representations (that can be finite or not) to propositional representations that can, in turn, incorporate more fine-grained non-probabilistic structure than mere beliefs (entrenchment or some notion of plausibility, for example).

The traditional approach derived from the seminal work of Carlos Alchourrón, Peter Gärdenfors and David Makinson (AGM) opted for a sentential representation. They were ambivalent between using finite representations (usually called belief bases) and using more structured representations of belief (theories in logical parlance or belief sets). Isaac Levi proposed a philosophical justification for the use of theories: they are supposed to represent not belief but commitments to belief.

So, how should one go about contracting a sentence A from a theory K? Merely deleting A will not do, given that the sentence might be entailed by other sentences in K. The AGM solution to the problem is simple. Enter the *remainder set* of K with A (denoted $K\perp A$). This is the set of all maximal subsets of K that fail to entail A. One then makes a selection from the members of this set via the use of a selection function γ. This gives us $\gamma(K\perp A)$. Finally one takes the intersection of the resulting set. This is usually called a *partial meet* contraction. Of course there are many possible selection functions that can be used in the previous account. AGM would in turn require that the selection can be rationalized in the usual sense common in rational choice, i.e. by assuming that the selection function selects the *best* elements of $K\perp A$ (with respect to an underlying transitive preference relation). Then AGM manage to offer a set of rationality postulates that completely characterizes this operation. This is the main result presented in the article reprinted here. This article single handedly created an entire field of research by combining ideas from rational choice and classical proof techniques from philosophical logic.

If one reads the original AGM article carefully one sees that much of the inspiration from the use of selection functions in their model comes from the seminal work of Amartya Sen in rational choice. Exploiting results from the theory of choice Hans Rott (and previously Sten Lindström) systematically studies the relationship between functional constraints placed on selection functions and postulates of belief change. Among other things, Rott shows that certain functional constraints placed on propositional selection functions correspond in a one-to-one fashion to postulates of belief change. Rott's results forge a useful bridge between the mathematical theories of belief change and rational choice.

Still one might inquire why the feasible set from which one chooses rationally via a selection function should be limited to $K\perp A$. One common defense of this idea is in terms of minimizing information loss. The elements of $K\perp A$, usually

called maxi-choice sets, do satisfy this requirement. But then if they are optimal until this point of view why not to use a selection function that picks singletons from $K\perp A$? AGM showed that this type of contraction is badly behaved. So is the opposite idea of taking directly the intersection of the entire $K\perp A$. So, partial meet appears as an Aristotelian middle ground that happens to satisfy a set of intuitive postulates. Or so argued AGM. Nevertheless the subsequent discussion focused on some controversial AGM postulates like *recovery* (requiring that if one contracts K with A and then adds A to the result of this contraction one returns to K). There are many putative counterexamples to recovery and this generated the interest in defining notions of contraction that fail to satisfy recovery. Isaac Levi is a well-known defender of this line of thought and in his article he characterizes a notion of contractions that does fail to satisfy recovery. The central idea he proposes is that what is minimized in contraction is not information loss but loss of *informational value*. The notion of information value is some sort of epistemic utility obeying basic structural postulates like:

(Weak Monotony) If $X \subset Y$, then $V(X) \leq V(Y)$.

This is an intuitive principle that makes permissible that two sets carry equal informational value even when one the sets carries more information than the other. The additional information might not be valuable at all and therefore the level of informational value of the larger set might remain equal to the informational value of the smaller set. What other constraints one should impose on information value? In the article reprinted here Levi presents a very specific form of information value that he uses to characterize a particular notion of withdrawal (some rational notion of contraction where recovery fails) that he calls *mild contraction*. Rott and Pagnucco offered an alternative model of the same notion that they call *severe withdrawal*. It is clear that when epistemic utility satisfies the constraints proposed by Levi this particular form of contraction obtains. What seems to be missing is a pre-systematic explanation of why epistemic utility should satisfy these constraints or a justification of some controversial properties of severe withdrawal (like the postulate of *antitony*). It is A true thought that the introduction of epistemic utility in models of belief change opens up an insightful research strategy that at the moment remains relatively unexplored.

Sven Ove Hansson offers another account of contraction that fails to obey recovery. Nevertheless he arrives at this conclusion in a completely different way. In fact, Hansson is one of the most prominent defenders of finite models of belief in terms of belief bases (finite sets of sentences that are one of the possible axiomatic bases of a given theory). It is easy to characterize a version of partial meet contraction for bases by using the obvious variant of the definition used for theories. Then one can proceed as follows: an operation of contraction on a belief set K is generated by a partial meet base contraction if and only if there is a belief base B for K and an operator $\sim$ of partial meet contraction for B such that the contraction of K with A yields the logical consequences of $(B \sim A)$ for all sentences A in the underlying language. Hansson shows that if an operation on a belief set is generated by some partial meet base contraction, then it satisfies the classical AGM postulates

for contraction except recovery. In addition the operation satisfies other postulates encoding a specific notion of *conservativity*.

The article by Spohn articulates an important epistemological idea, namely that one should focus on changes of entire *epistemic states* endowed with more structure than mere belief. This approach, in a more general setting, is also independently pursued by Adnan Darwiche and Judea Pearl in Darwiche and Pearl (1996). Spohn focuses on a particular type of epistemic state that now is usually called a *ranking function*. Roughly a ranking function is a function from the set of propositions (= sets of possible worlds) to the set of natural, real, or ordinal numbers, similar to a probability measure. Epistemologically one can see such functions as numerical (but non-probabilistic) representations of a notion of *plausibility*. In the presence of a new input the current ranking is mapped to a new ranking incorporating the incoming information (in revision). This is an ideal setting to study the structure of iterated changes of view and as a matter of fact both articles offer the best existing articulation of principles regulating iterated change. This is an important area of research in this field that still remains relatively open.

Suggested Further Reading

There are a number of recent surveys and books that complement the reprinted papers appearing here. Regarding surveys the two most recent surveys are: Logic of Belief Revision, in *Stanford Encyclopedia of Philosophy,* 2006, by Sven Ove Hansson; and: Belief Revision in *The Continuum Companion to Philosophical Logic*, (eds.) L. Hornsten and R. Pettigrew, by Horacio Arlo-Costa and Paul Pedersen. These surveys contain references to previous surveys in the field. A classic book in this area that continues to be useful is Peter Gärdenfors's monograph: *Knowledge in Flux: Modeling the Dynamic of Epistemic States*, College Publications (June 2, 2008). A very useful textbook presentation of some of the main results in the theory of belief change is: *A Textbook of Belief Dynamics: Theory Change and Database Updating*, Springer 2010, by Sven Ove Hansson. The book focuses mainly on syntactic presentations of belief change and it contains a very detailed presentation of belief base updating. Some more recent topics like iterated belief change are not treated in detail though.

Decision theoretic foundations for belief change are provided in various books by Hans Rott and Isaac Levi (independently). A book-length argument articulating Rott's account (and extending the content of the article reprinted here) appears in: *Change, Choice and Inference: A Study of Belief Revision and Non-monotonic Reasoning*, Oxford Logic Guides, 2001. Some challenges to this type of foundational strategy are considered by Arlo-Costa and Pedersen in: "Social Norms, Rational Choice and Belief Change," in *Belief Revision Meets Philosophy of Science*, (eds.) E.J. Olsson and S. Enqvist, Springer, 2011. Isaac Levi has also published various essays where he presents decision theoretic foundations for belief change (but his account is rather different than Rott's). The most recent book presenting Levi's current views about belief change is: *Mild Contraction: Evaluating Loss of Information Due to Loss of Belief*, Oxford, 2004. Further references to his work can be found in this book.

The previous accounts tried to justify principles of belief change in the broader context of Bayesian or neo-Bayesian theory. An almost orthogonal view consists in deriving principles of belief change by taking some form of formal learning theory as an epistemological primitive. While all the previous accounts focused on justifying the next step of inquiry (or a finite and proximate sequence of steps) this second strategy focuses on selecting belief change methods capable of learning the truth in the long run. One important paper in this tradition is Kevin Kelly's: Iterated

Belief Revision, Reliability, and Inductive Amnesia, *Erkenntnis*, 50, 1998 pp. 11–58. Daniel Osherson and Eric Martin present a similarly motivated account that nevertheless is formally quite different from Kelly's theory in: *Elements of Scientific Inquiry*, MIT, 1998.

There are various attempts to extend the theory of belief revision to the multi-agent case and to present a theory of belief change as some form of *dynamic epistemic logic*. The idea in this case is to use traditional formal tools in epistemic logic to represent the process of belief change. Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi have recently published a textbook with some basic results in this area: *Dynamic Epistemic Logic*, Springer, 2011. Krister Segerberg has developed his own brand of dynamic doxastic logic in a series of articles since at least the mid 1990's. One recent paper including rather comprehensive results in this area is: "Some Completeness Theorems in the Dynamic Doxastic Logic of Iterated Belief Revision," *Review of Symbolic Logic*, 3(2):228–246, 2010.

The notion of *relevance* is quite central for a representation of belief and belief change. In a Bayesian setting there are standard ways of articulating relevance. But there is recent work that has used proof theoretic techniques to deal with relevance rather than probability theory. Rohit Parikh initiated this area of research with an article published in 1999: Beliefs, belief revision, and splitting languages, *Logic, language, and computation* (Stanford, California) (Lawrence Moss, Jonathan Ginzburg, and Maarten de Rijke, editors), vol. 2, CSLI Publications, pp. 266–278. Recently David Makinson has contributed as well an important article in collaboration with George Kourousias,: Parallel interpolation, splitting, and relevance in belief change, *Journal of Symbolic Logic* 72 September 2007 994-1002. This article contains a detailed bibliography of recent work in this area.

One recent paper including rather comprehensive results in this area is: "Some completeness theorems in the dynamic doxastic logic of iterated belief revision," *Review of Symbolic Logic, 3*, 02, 2010. For more on iterated belief revision please refer to: Darwiche and Pearl (Darwiche, A., & Pearl, J. (1996). On the logic of iterated belief revision. *Artificial Intelligence, 89*, 1–29) appears in: *Change, choice and inference: A study of belief revision and non-monotonic reasoning*, Oxford Logic Guides, 2001.

And there is also more to be found in Pagnucco and Rott (Pagnucco, M., & Rott, H. (1999). Severe withdrawal – and recovery. *Journal of Philosophical Logic, 28*, 501–547. See publisher's "Erratum" (2000), *Journal of Philosophical Logic, 29*, 121) and Lindström (Lindström, S. (1991). A semantic approach to nonmonotonic reasoning: Inference operations and choice. *Uppsala Prints and Preprints in Philosophy*, no. 6/1991, University of Uppsala).

Chapter 13
On the Logic of Theory Change: Partial Meet Contraction and Revision Functions

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson

1 Background

The simplest and best known form of theory change is *expansion*, where a new proposition (axiom), hopefully consistent with a given theory A, is set-theoretically added to A, and this expanded set is then closed under logical consequence. There are, however, other kinds of theory change, the logic of which is less well understood. One form is theory *contraction*, where a proposition x, which was earlier in a theory A, is rejected. When A is a code of norms, this process is known among legal theorists as the *derogation* of x from A. The central problem is to determine which propositions should be rejected along with x so that the contracted theory will be closed under logical consequence. Another kind of change is *revision*, where a proposition x, inconsistent with a given theory A, is added to A under the requirement that the revised theory be consistent and closed under logical consequence. In normative contexts this kind of change is also known as *amendment*.

Carlos E. Alchourrón died on 13 January 1996.

This paper was written while the third author was on leave from UNESCO. The contents are the responsibility of the authors and not of the institutions.

C.E. Alchourrón (deceased)
Universidad de Buenos Aires, Buenos Aires, Argentina

P. Gärdenfors
Lund University, Lund, Sweden

D. Makinson
London School of Economics, London, UK

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_13

A basic formal problem for the processes of contraction and revision is to give a characterization of ideal forms of such change. In (Gärdenfors 1978) and (Gärdenfors 1982), Gärdenfors developed postulates of a more or less equational nature to capture the basic properties of these processes. It was also argued there that the process of revision can be reduced to that of contraction via the so-called Levi identity: if $A \dot{-} x$ denotes the contraction of A by x, then the revision of A by x, denoted $A \dot{+} x$, can be defined as $\mathrm{Cn}((A \dot{-} \neg x) \cup \{x\})$, where Cn is a given consequence operation.

In (Alchourron and Makinson 1982b), Alchourrón and Makinson tried to give a more explicit construction of the contraction process, and hence also of the revision process via the Levi identity. Their basic idea was to choose $A \dot{-} x$ as a *maximal* subset of A that fails to imply x. Contraction functions defined in this way were called "choice contractions" in (Alchourron and Makinson 1982b), but will here be more graphically referred to as "maxichoice contractions".

As was observed in (Alchourron and Makinson 1982b), the maxichoice functions have, however, some rather disconcerting properties. In particular, maxichoice revision $\dot{+}$, defined from maxichoice contraction as above, has the property that for every theory A, whether complete or not, $A \dot{+} x$ will be complete whenever x is a proposition inconsistent with A. Underlying this is the fact, also noted in (Alchourron and Makinson 1982b), that when A is a theory with $x \in A$, then for every proposition y, either $(x \vee y) \in A \dot{-} x$ or $(x \vee \neg y) \in A \dot{-} x$, where $\dot{-}$ is maxichoice contraction. The significance of these formal results is discussed briefly in (Alchourron and Makinson 1982b), and in more detail in Gärdenfors (1984) and Makinson (1985).

The "inflation properties" that ensue from applying the maxichoice operations bring out the interest of looking at other formal operations that yield smaller sets as values. In this paper, we will start out from the assumption that there is a selection function γ that picks out a class of the "most important" maximal subsets of A that fail to imply x. The contraction $A \dot{-} x$ is then defined as the intersection of all the maximal subsets selected by γ. Functions defined in this way will be called *partial meet contraction* functions, and their corresponding revision functions will be called *partial meet revision* functions. It will be shown that they satisfy Gärdenfors' postulates, and indeed provide a representation theorem for those postulates. When constrained in suitable ways, by relations or, more restrictedly, by transitive relations, they also satisfy his "supplementary postulates", and provide another representation theorem for the entire collection of "basic" plus "supplementary" postulates.

Acquaintance with (Makinson 1985) will help the reader with overall perspective, but it is not necessary for technical details.

Some background terminology and notation: By a *consequence operation* we mean, as is customary, an operation Cn that takes sets of propositions to sets of propositions, such that three conditions are satisfied, for any sets X and Y of propositions: $X \subseteq \mathrm{Cn}(X)$, $\mathrm{Cn}(X) = \mathrm{Cn}(\mathrm{Cn}(X))$, and $\mathrm{Cn}(X) \subseteq \mathrm{Cn}(Y)$ whenever $X \subseteq Y$. To simplify notation, we write $\mathrm{Cn}(x)$ for $\mathrm{Cn}(\{x\})$, where x is any individual proposition, and we also sometimes write $y \in \mathrm{Cn}(X)$ as $X \vdash y$. By a *theory*, we

mean, as is customary, a set A of propositions that is closed under Cn; that is, such that $A = \mathrm{Cn}(A)$, or, equivalently, such that $A = \mathrm{Cn}(B)$ for some set B of propositions. As in (Alchourron and Makinson 1982b), we assume that Cn includes classical tautological implication, is compact (that is, $y \in \mathrm{Cn}(X')$ for some finite subset X' of X whenever $y \in \mathrm{Cn}(X)$), and satisfies the rule of "introduction of disjunctions in the premises" (that is, $y \in \mathrm{Cn}(X \cup \{x_1 \vee x_2\})$ whenever $y \in \mathrm{Cn}(X \cup \{x_1\})$ and $y \in \mathrm{Cn}(X \cup \{x_2\})$). We say that a set X of propositions is *consistent* (modulo Cn) iff for no proposition y do we have $y \,\&\, \neg y \in \mathrm{Cn}(X)$.

2 Partial Meet Contraction

Let Cn be any consequence operation over a language, satisfying the conditions mentioned at the end of the preceding section, and let A be any set of propositions. As in (Alchourron and Makinson 1982a) and (Alchourron and Makinson 1982b), we define $A \perp x$ to be the set of all maximal subsets B of A such that $B \nvdash x$. The maxichoice contraction functions $\dot{-}$ studied in (Alchourron and Makinson 1982b) put $A \dot{-} x$ to be an arbitrary element of $A \perp x$ whenever the latter is nonempty, and to be A itself in the limiting case that $A \perp x$ is empty. In the search for suitable functions with smaller values, it is tempting to try the operation $A \sim x$ defined as $\cap(A \perp x)$ when $A \perp x$ is nonempty, and as A itself in the limiting case that $A \perp x$ is empty. But as shown in Observation 2.1 of (Alchourron and Makinson 1982b), this set is in general far too *small*. In particular, when A is a theory with $x \in A$, then $A \sim x = A \cap \mathrm{Cn}(\neg x)$. In other words, the only propositions left in $A \sim x$ when A is a theory containing x are those which are already consequences of $\neg x$ considered alone. And thus, as noted in Observation 2.2 of (Alchourron and Makinson 1982b), if revision is introduced as usual via the Levi identity as $\mathrm{Cn}((A \sim \neg x) \cup \{x\})$, it reduces to $\mathrm{Cn}((A \cap \mathrm{Cn}(x)) \cup \{x\}) = \mathrm{Cn}(x)$, for any theory A and proposition x inconsistent with A. In other words, if we revise a theory A in *this* way to bring in a proposition x inconsistent with A, we get no more than the set of consequences of x considered alone—a set which is far too small in general to represent the result of an intuitive process of revision of A so as to bring in x.

Nevertheless, the operation of *meet contraction*, as we shall call $\sim$, is very useful as a point of reference. It serves as a natural *lower bound* on any reasonable contraction operation: any contraction operation $\dot{-}$ worthy of the name should surely have $A \sim x \subseteq A \dot{-} x$ for all A, x, and a function $\dot{-}$ satisfying this condition for a given A will be called *bounded over* A.

Following this lead, let A be any set of propositions and let γ be any function such that for every proposition x, $\gamma(A \perp x)$ is a nonempty subset of $A \perp x$, if the latter is nonempty, and $\gamma(A \perp x) = \{A\}$ in the limiting case that $A \perp x$ is empty. We call such a function a *selection function for* A. Then the operation $\dot{-}$ defined by putting $A \dot{-} x = \cap\gamma(A \perp x)$ for all x is called the *partial meet contraction over* A *determined by* γ. The intuitive idea is that the selection function γ picks out those elements in $A \perp x$ which are "most important" (for a discussion of this notion cf. Gärdenfors (1984))

and then the contraction $A \dot{-} x$ contains the propositions which are common to the selected elements of $A \perp x$. *Partial meet revision* is defined via the Levi identity as $A \dot{+} x = \mathrm{Cn}((A \dot{-} \neg x) \cup \{x\})$. Note that the identity of $A \dot{-} x$ and $A \dot{+} x$ depends on the choice function γ, as well, of course, as on the underlying consequence operation Cn. Note also that the concept of partial meet contraction includes, as special cases, those of maxichoice contraction and (full) meet contraction. The former is partial meet contraction with $\gamma(A \perp x)$ a singleton; the latter is partial meet contraction with $\gamma(A \perp x)$ the entire set $A \perp x$. We use the same symbols $\dot{-}$ and $\dot{+}$ here as for the maxichoice operations in (Alchourron and Makinson 1982b); this should not cause any confusion.

Our first task is to show that all partial meet contraction and revision functions satisfy Gärdenfors' postulates for contraction and revision. We recall (cf. (Alchourron and Makinson 1982b) and (Makinson 1985)) that these postulates may conveniently be formulated as follows:

($\dot{-}1$) $A \dot{-} x$ is a theory whenever A is a theory (closure).
($\dot{-}2$) $A \dot{-} x \subseteq A$ (inclusion).
($\dot{-}3$) If $x \notin \mathrm{Cn}(A)$, then $A \dot{-} x = A$ (vacuity).
($\dot{-}4$) If $x \notin \mathrm{Cn}(\varnothing)$, then $x \notin \mathrm{Cn}(A \dot{-} x)$ (success).
($\dot{-}5$) If $\mathrm{Cn}(x) = \mathrm{Cn}(y)$, then $A \dot{-} x = A \dot{-} y$ (preservation).
($\dot{-}6$) $A \subseteq \mathrm{Cn}((A \dot{-} x) \cup \{x\})$ whenever A is a theory (recovery).

The Gärdenfors postulates for revision may likewise be conveniently formulated as follows:

($\dot{+}1$) $A \dot{+} x$ is always a theory.
($\dot{+}2$) $x \in A + x$.
($\dot{+}3$) If $\neg x \notin \mathrm{Cn}(A)$, then $A \dot{+} x = \mathrm{Cn}\,(A \cup \{x\})$.
($\dot{+}4$) If $\neg x \notin \mathrm{Cn}(\varnothing)$, then $A \dot{+} x$ is consistent under Cn.
($\dot{+}5$) If $\mathrm{Cn}(x) = \mathrm{Cn}(y)$, then $A \dot{+} x = A \dot{+} y$.
($\dot{+}6$) $(A \dot{+} x) \cap A = A \dot{-} \neg x$, whenever A is a theory.

Our first lemma tells us that even the very weak operation of (full) meet contraction satisfies recovery.

Lemma 2.1 *Let A be any theory. Then $A \subseteq \mathrm{Cn}((A \sim x) \cup \{x\})$.*

Proof In the limiting case that $x \notin A$ we have $A \sim x = A$ and we are done. Suppose $x \in A$. Then, by Observation 2.1 of (Alchourron and Makinson 1982b), we have $A \sim x = A \cap \mathrm{Cn}(\neg x)$ so it will suffice to show $A \subseteq \mathrm{Cn}((A \cap \mathrm{Cn}(\neg x)) \cup \{x\})$. Let $a \in A$. Then since A is a theory, $\neg x \vee a \in A$. Also $\neg x \vee a \in \mathrm{Cn}(\neg x)$, so $\neg x \vee a \in A \cap \mathrm{Cn}(\neg x)$, so since Cn includes tautological implication, $a \in \mathrm{Cn}((A \cap \mathrm{Cn}(\neg x)) \cup \{x\})$. □

Corollary 2.2 *Let $\dot{-}$ be any function on pairs A, x. Let A be any theory. If $\dot{-}$ is bounded over A, then $\dot{-}$ satisfies recovery over A.*

Observation 2.3 *Every partial meet contraction function $\dot{-}$ satisfies the Gärdenfors postulates for contraction, and its associated partial meet revision function satisfies the Gärdenfors postulates for revision.*

Proof It is easy to show (cf. (Gärdenfors 1978) and (Gärdenfors 1982)) that the postulates for revision can all be derived from those for contraction via the Levi identity. So we need only verify the postulates for contraction. Closure holds, because when A is a theory, so too is each $B \in A \perp x$, and the intersection of theories is a theory; inclusion is immediate; vacuity holds because when $x \notin \mathrm{Cn}(A)$ then $A \perp x = \{A\}$ so $\gamma(A \perp x) = \{A\}$; success holds because when $x \notin \mathrm{Cn}(\varnothing)$ then by compactness, as noted in Observation 2.2 of (Alchourron and Makinson 1982a), $A \perp x$ is nonempty and so $A \dot{-} x = \cap\gamma(A \perp x) \nvdash x$; and preservation holds because the choice function is defined on families $A \perp x$ rather than simply on pairs A, x, so that when $\mathrm{Cn}(x) = \mathrm{Cn}(y)$ we have $A \perp x = A \perp y$, so that $\gamma\{A \perp x) = \gamma\ (A \perp y)$. Finally, partial meet contraction is clearly bounded over any set A, and so by Corollary 2.2 satisfies recovery. □

In fact, we can also prove a converse to Observation 2.3, and show that for theories, the Gärdenfors postulates for contraction *fully characterize* the class of partial meet contraction functions. To do this we first establish a useful general lemma related to 7.2 of (Alchourron and Makinson 1982b).

Lemma 2.4 *Let A be a theory and x a proposition. If $B \in A \perp x$, then $B \in A \perp y$ for all $y \in A$ such that $B \nvdash y$.*

Proof Suppose $B \in A \perp x$ and $B \nvdash y$, $y \in A$. To show that $B \in A \perp y$ it will suffice to show that whenever $B \subset B' \subseteq A$, then $B' \vdash y$. Let $B \subset B' \subseteq A$. Since $B \in A \perp x$ we have $B' \vdash x$. But also, since $B \in A \perp x$, $A \perp x$ is nonempty, so $A \sim x = \cap(A \perp x) \subseteq B$; so, using Lemma 2.1, $A \subseteq \mathrm{Cn}(B \cup \{x\}) \subseteq \mathrm{Cn}(B' \cup \{x\}) = \mathrm{Cn}(B')$, so since $y \in A$ we have $B' \vdash$ y. □

Observation 2.5 *Let $\dot{-}$ be a function defined for sets A of propositions and propositions x. For every theory A, $\dot{-}$ is a partial meet contraction operation over A iff $\dot{-}$ satisfies the Gärdenfors postulates ($\dot{-}$1) – ($\dot{-}$6) for contraction over A.*

Proof We have left to right by Observation 2.3. For the converse, suppose that $\dot{-}$ satisfies the Gärdenfors postulates over A. To show that $\dot{-}$ is a partial meet contraction operation, it will suffice to find a function such that:

(i) $\gamma(A \perp x) = \{A\}$ in the limiting case that $A \perp x$ is empty,
(ii) $\gamma(A \perp x)$ is a nonempty subset of $A \perp x$ when $A \perp x$ is nonempty, and
(iii) $A \dot{-} x = \bigcap\gamma(A \perp x)$.

Put $\gamma(A \perp x)$ to be $\{A\}$ when $A \perp x$ is empty, and to be $\{B \in A \perp x\colon A \dot{-} x \subseteq B\}$ otherwise. Then (i) holds immediately. When $A \perp x$ is nonempty, then $x \notin \mathrm{Cn}(\varnothing)$ so by the postulate of success $A \dot{-} x \nvdash x$, so, using compactness, $\gamma(A \perp x)$ is nonempty, and clearly $\gamma(A \perp x) \subseteq A \perp x$, so (ii) also holds. For (iii) we have the inclusion $A \dot{-} x \subseteq \bigcap\gamma(\mathrm{A} \perp x)$ immediately from the definition of γ. So it remains only to show that $\bigcap\gamma(A \perp x) \subseteq A \dot{-} x$.

In the case that $x \notin A$ we have by the postulate of vacuity that $A \dot{-} x = A$, so the desired conclusion holds trivially. Suppose then that $x \in A$, and suppose $a \notin A \dot{-} x$; we want to show that $a \notin \cap\gamma(A \perp x)$. In the case $a \notin A$, this holds trivially, so we

suppose that $a \in A$. We need to find a $B \in A \perp x$ with $A \dot{-} x \subseteq B$ and $a \notin B$. Since $\dot{-}$ satisfies the postulate of recovery, and $a \in A$, we have $(A \dot{-} x) \cup \{x\} \vdash a$. But, by hypothesis, $a \notin A \dot{-} x = \mathrm{Cn}(A \dot{-} x)$ by the postulate of closure, so since Cn includes tautological implication and satisfies disjunction of premises, $(A \dot{-} x) \cup \{\neg x\} \nvdash a$, so $A \dot{-} x \nvdash x \vee a$. Hence by compactness there is a $B \in A \perp (x \vee a)$ with $A \dot{-} x \subseteq B$. Since $B \in A \perp (x \vee a)$ we have $B \nvdash x \vee$ a, so $a \notin B$. And also since $B \nvdash x \vee a$ we have $B \nvdash x$, so, by Lemma 2.4, and the hypothesis that $x \in A$, we have $B \in A \perp x$, and the proof is complete. □

A corollary of Observation 2.5 is that whenever $\dot{-}$ satisfies the Gärdenfors postulates for contraction over a theory A, then it is bounded over A. However, this consequence can also be obtained, under slightly weaker conditions, by a more direct argument. We first note the following partial converse of Lemma 2.1.

Lemma 2.6 *Let A be any theory. Then for every set B and every $x \in A$, if $A \subseteq \mathrm{Cn}(B \cup \{x\})$, then $A \sim x \subseteq \mathrm{Cn}(B)$.*

Proof Suppose $x \in A$, $A \subseteq \mathrm{Cn}(B \cup \{x\})$, and $a \in A \sim x$; we want to show that $a \in \mathrm{Cn}(B)$. Since A is a theory and $x \in A$ we have $A \sim x = \mathrm{Cn}(\neg x) \cap A$ by Observation 2.1 of (Alchourron and Makinson 1982b); so $\neg x \vdash a$, so $B \cup \{\neg x\} \vdash a$. But also since $a \in A \sim x \subseteq A \subseteq \mathrm{Cn}(B \cup \{x\})$ we have $B \cup \{x\} \vdash a$, so by disjunction of premises and the fact that Cn includes tautological implication, we have $a \in \mathrm{Cn}(B)$. □

Observation 2.7 *Let $\dot{-}$ be any function on pairs A, x. Let A be a theory. If $\dot{-}$ satisfies closure, vacuity and recovery over A, then $\dot{-}$ is bounded over A.*

Proof Suppose $\dot{-}$ satisfies closure, vacuity and recovery over A. Let x be any proposition; we need to show $A \sim x \subseteq A \dot{-} x$. In the case $x \in A$ we have trivially $A \sim x = A \dot{-} x$ by vacuity. In the case $x \in A$ we have $A \subseteq \mathrm{Cn}((A \dot{-} x) \cup \{x\})$ by recovery, so $A \sim x \subseteq \mathrm{Cn}(A \dot{-} x) = A \dot{-} x$ by Lemma 2.6 and closure. □

3 Supplementary Postulates for Contraction and Revision

Gärdenfors (1984) has suggested that revision should also satisfy two further "supplementary postulates", namely:

($\dot{+}7$) $A \dot{+} (x \,\&\, y) \subseteq \mathrm{Cn}((A \dot{+} x) \cup \{y\})$ for any theory A, and its conditional converse:
($\dot{+}8$) $\mathrm{Cn}((A \dot{+} x) \cup \{y\}) \subseteq A \dot{+} (x \,\&\, y)$ for any theory A, provided that $\neg y \notin A \dot{+} x$.

Given the presence of the postulates ($\dot{-}1$) – ($\dot{-}6$) and ($\dot{+}1$) – ($\dot{+}6$), these two supplementary postulates for $\dot{+}$ can be shown to be equivalent to various conditions on $\dot{-}$. Some such conditions are given in (Gärdenfors 1984); these can however be simplified, and one particularly simple pair, equivalent respectively to ($\dot{+}7$) and ($\dot{+}8$), are:

($\dot{-}7$) $(A \dot{-} x) \cap (A \dot{-} y) \subseteq A \dot{-} (x \,\&\, y)$ for any theory A.
($\dot{-}8$) $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x$ whenever $x \notin A \dot{-} (x \,\&\, y)$, for any theory A.

Observation 3.1 *Let $\dot{-}$ be any partial meet contraction operation over a theory A. Then it satisfies* ($\dot{-}7$) *iff it satisfies* ($\dot{+}7$).

Proof We recall that $\dot{+}$ is defined by the Levi identity $A \dot{+} x = \mathrm{Cn}((A \dot{-} \neg x) \cup \{x\})$. Let A be any theory and suppose that ($\dot{-}7$) holds for all x and y. We want to show that ($\dot{+}7$) holds for all x and y. Let

$$w \in A \dot{+} (x\&y) = \mathrm{Cn}\left((A \dot{-} \neg (x\&y)) \cup \{x\&y\}\right).$$

We need to show that

$$\begin{aligned} w \in \mathrm{Cn}\left(\left(A \dot{+} x\right) \cup \{y\}\right) &= \mathrm{Cn}\left(\mathrm{Cn}\left((A \dot{-} \neg x) \cup \{x\}\right) \cup \{y\}\right) \\ &= \mathrm{Cn}\left((A \dot{-} \neg x) \cup \{x\&y\}\right) \end{aligned}$$

by general properties of consequence operations. Noting that

$$\mathrm{Cn}(\neg x) = \mathrm{Cn}\left(\neg (x\&y) \,\&\, (\neg x \vee y)\right),$$

it will suffice by condition ($\dot{-}7$) to show that

$$w \in \mathrm{Cn}\left((A \dot{-} \neg (x\&y)) \cup \{x\&y\}\right) \text{ and } w \in \mathrm{Cn}\left((A \dot{-} (\neg x \vee y)) \cup \{x\&y\}\right).$$

But the former is given by hypothesis, so we need only verify the latter. Now by the former, we have $w \in \mathrm{Cn}(A \cup \{x \,\&\, y\})$, so it will suffice to show that

$$A \cup \{x\&y\} \subseteq \mathrm{Cn}\left((A \dot{-} (\neg x \vee y)) \cup \{x\&y\}\right).$$

But clearly $x \,\&\, y \in \mathrm{RHS}$, and moreover since $x \,\&\, y \vdash y \vdash \neg x \vee y$ we have by recovery that $A \subseteq \mathrm{RHS}$, and we are done.

For the converse, suppose that ($\dot{+}7$) holds for all x, y. Let $a \in (A \dot{-} x) \cap (A \dot{-} y)$; we need to show that $a \in A \dot{-} (x \,\&\, y)$. Noting that

$$\mathrm{Cn}(x) = \mathrm{Cn}\left(\neg \left((\neg x \vee \neg y) \,\&\, \neg x\right)\right),$$

we have

$$\begin{aligned} a \in A \dot{-} \neg \left((\neg x \vee \neg y) \,\&\, \neg x\right) &\subseteq A \dot{+} \left((\neg x \vee \neg y) \,\&\, \neg x\right) \\ &\subseteq \mathrm{Cn}\left(\left(A \dot{+} (\neg x \vee \neg y)\right) \cup \{\neg x\}\right). \end{aligned}$$

A similar reasoning gives us also $a \in \mathrm{Cn}((A \dot{+} (\neg x \vee \neg y)) \cup \{\neg y\})$. So applying disjunction of premises and the fact that Cn includes tautological implication, we have

$$a \in \mathrm{Cn}\left(A \dot{+} (\neg x \vee \neg y)\right) = A \dot{+} (\neg x \vee \neg y) = Cn\left((A \dot{-} (x\&y)) \cup \{\neg (x\&y)\}\right).$$

But by recovery we also have $a \in \mathrm{Cn}((A \dot{-} (x \,\&\, y)) \cup \{x \,\&\, y\})$, so, again using disjunction of premises,

$$a \in \mathrm{Cn}\,(A \dot{-} (x \& y)) = A \dot{-} (x \& y)$$

by closure, and we are done. □

Observation 3.2 *Let $\dot{-}$ be any partial meet contraction function over a theory A. Then it satisfies* ($\dot{-}8$) *iff it satisfies* ($\dot{+}8$).

Proof Let A be a theory and suppose that ($\dot{-}8$) holds for all x and y. We want to show that ($\dot{+}8$) holds for all x and y. Noting that $\mathrm{Cn}(\neg x) = \mathrm{Cn}((\neg x \vee \neg y) \,\&\, \neg x)$ we have $A \dot{-} \neg x = A \dot{-} ((\neg x \vee \neg y) \,\&\, \neg x)$. But also, supposing for ($\dot{+}8$) that $\neg y \notin A \dot{+} x = \mathrm{Cn}((A \dot{-} \neg x) \cup \{x\})$, we have $\neg x \vee \neg y \notin A \dot{-} \neg x$. We may thus apply ($\dot{-}8$) to get

$$A \dot{-} \neg x = A \dot{-} ((\neg x \vee \neg y) \,\&\, \neg x) \subseteq A \dot{-} (\neg x \vee \neg y) = A \dot{-} \neg\, (x \& y)\,.$$

This inclusion justifies the inclusion in the following chain, whose other steps are trivial:

$$\begin{aligned} \mathrm{Cn}\left(\left(A \dot{+} x\right) \cup \{y\}\right) &= \mathrm{Cn}\,(\mathrm{Cn}\,((A \dot{-} \neg x) \cup \{x\}) \cup \{y\}) \\ &= \mathrm{Cn}\,((A \dot{-} \neg x) \cup \{x \& y\}) \subseteq \mathrm{Cn}\,((A \dot{-} \neg\, (x \& y)) \cup \{x \& y\}) \\ &= A \dot{+} (x \& y)\,. \end{aligned}$$

For the converse, suppose ($\dot{+}8$) holds for all x and y, and suppose $x \notin A \dot{-} (x \,\&\, y)$. Then clearly

$$x \notin \mathrm{Cn}\,(A \dot{-} (x \& y) \cup \{\neg x \vee \neg y\}) = A \dot{+} \neg\, (x \& y)\,,$$

so we may apply ($\dot{+}8$) to get

$$\begin{aligned} \mathrm{Cn}\left(\left(A \dot{+} \neg\, (x \& y)\right) \cup \{\neg x\}\right) \subseteq A \dot{+} (\neg\, (x \& y) \,\&\, \neg x) &= A \dot{+} \neg x \\ &= \mathrm{Cn}\,((A \dot{-} x) \cup \{\neg x\})\,. \end{aligned}$$

Thus, since $A \dot{-} (x \,\&\, y)$ is included in the leftmost term of this series, we have

$$A \dot{-} (x \& y) \subseteq \mathrm{Cn}\,((A \dot{-} x) \cup \{\neg x\})\,.$$

But using recovery we also have $A \dot{-} (x \,\&\, y) \subseteq A \subseteq \mathrm{Cn}((A \dot{-} x) \cup \{x\})$, so by disjunction of premises and the fact that Cn includes tautological implication, we have $A \dot{-} (x \,\&\, y) \subseteq \mathrm{Cn}(A \dot{-} x) = A \dot{-} x$ by closure, as desired. □

We end this section with some further observations on the powers of ($\dot{-}7$) and ($\dot{-}8$). Now postulate ($\dot{-}7$) does not tell us that $A \dot{-} x$ and $A \dot{-} y$, considered

separately, are included in $A \dot{-} (x \,\&\, y)$. But it goes close to it, for it does yield the following "partial antitony" property.

Observation 3.3 *Let $\dot{-}$ be any partial meet contraction function over a theory A. Then $\dot{-}$ satisfies* ($\dot{-}$7) *iff it satisfies the condition*

($\dot{-}$P) $(A \dot{-} x) \cap \mathrm{Cn}(x) \subseteq A \dot{-} (x \,\&\, y)$ *for all x and y.*

Proof Suppose ($\dot{-}$7) is satisfied. Suppose $w \in A \dot{-} x$ and $x \vdash w$; we want to show that $w \in A \dot{-} (x \,\&\, y)$. If $x \notin A$ or $y \notin A$, then trivially $A \dot{-} (x \,\&\, y) = A$, so $w \in A \dot{-} (x \,\&\, y)$. So suppose that $x \in A$ and $y \in A$. Now

$$A \dot{-} (x \& y) = A \dot{-} ((\neg x \vee y) \;\&\; x)\,,$$

so by ($\dot{-}$7) it will suffice to show that $w \in A \dot{-} (\neg x \vee y)$ and $w \in A \dot{-} x$. We have the latter by supposition. As for the former, recovery gives us $A \dot{-} (\neg x \vee y) \cup \{\neg x \vee y) \vdash x$, so $A \dot{-} (\neg x \vee y) \cup \{\neg x\} \vdash x$, so $A \dot{-} (\neg x \vee y) \vdash x \vdash w$, so $w \in A \dot{-} (\neg x \vee y)$.

For the converse, suppose ($\dot{-}$P) is satisfied, and suppose $w \in (A \dot{-} x) \cap (A \dot{-} y)$; we want to show that $w \in A \dot{-} (x \,\&\, y)$. Since $w \in A \dot{-} x$, we have $x \vee w \in A \dot{-} x$, and so since $x \vdash x \vee w$, ($\dot{-}$P) gives us $x \vee w \in A \dot{-} (x \,\&\, y)$. Similarly, $y \vee w \in A \dot{-} (x \,\&\, y)$. Hence $w \vee (x \,\&\, y) = (x \vee w) \,\&\, (y \vee w) \in A \dot{-} (x \,\&\, y)$. But by recovery, $A \dot{-} (x \,\&\, y) \cup \{x \,\&\, y\} \vdash w$, so $w \vee \neg(x \,\&\, y) \in A \dot{-} (x \,\&\, y)$. Putting these together gives us $w \in A \dot{-} (x \,\&\, y)$ as desired. □

Condition ($\dot{-}$8) is related to another condition, which we shall call the *covering condition*:

($\dot{-}$C) For any propositions x, y, $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x$ or $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} y$.

Observation 3.4 *Let $\dot{-}$ be any partial meet contraction function over a theory A. If $\dot{-}$ satisfies* ($\dot{-}$8) *over A, then it satisfies the covering condition* ($\dot{-}$C) *over A.*

Proof Let x and y be propositions. In the case $x \,\&\, y \in \mathrm{Cn}(\varnothing)$ we have, say, $x \in \mathrm{Cn}(\varnothing)$; so $A \dot{-} (x \,\&\, y) = A = A \dot{-} x$ and we are done. In the case $x \,\&\, y \notin \mathrm{Cn}(\varnothing)$, then by success we have $x \,\&\, y \notin A \dot{-} (x \,\&\, y)$, so either $x \notin A \dot{-} (x \,\&\, y)$ or $y \notin A \dot{-} (x \,\&\, y)$, so by ($\dot{-}$8) either $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x$ or $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} y$. □

However, the converse of Observation 3.4 fails. For as we shall show at the end of the next section, there is a theory, finite modulo Cn, with a partial meet contraction over A that satisfies the covering condition (and indeed also supplementary postulate ($\dot{-}$7)), but that does not satisfy ($\dot{-}$8). Using Observation 3.4, it is easy to show that when A is a theory and $\dot{-}$ satisfies postulates ($\dot{-}$1) – ($\dot{-}$6), then ($\dot{-}$8) can equivalently be formulated as $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x$ whenever $x \notin A \dot{-} y$.

In (Alchourron and Makinson 1982b), it was shown that whilst the *maxichoice* operations do not in general satisfy ($\dot{-}$7) and ($\dot{-}$8), they do so when constrained by a relational condition of "orderliness". Indeed, it was shown that for the maxichoice operations, the conditions ($\dot{+}$7), ($\dot{+}$8), and orderliness are mutually equivalent, and also equivalent to various other conditions. Now as we have just remarked, in the general context of partial meet contraction, ($\dot{-}$7) does not imply ($\dot{-}$8), and it can

also be shown by an example (briefly described at the end of next section) that the converse implication likewise fails. The question nevertheless remains whether there are relational constraints on the partial meet operations that correspond, perfectly or in part, to the supplementary postulates ($\dot{-}7$) and ($\dot{-}8$). That is the principal theme of the next section.

4 Partial Meet Contraction with Relational Constraints

Let A be a set of propositions and γ a selection function for A. We say that γ is *relational* over A iff there is a relation $\leq$ over 2^A such that for all $x \notin \mathrm{Cn}(\varnothing)$, $\leq$ *marks off* $\gamma(A \perp x)$ in the sense that the following identity, which we call the *marking off* identity, holds:

$$\gamma\,(A \perp x) = \left\{B \in A \perp x : B' \;\leq\; B \text{ for all } B' \in A \perp x\right\}.$$

Roughly speaking, γ is relational over A iff there is some relation that marks off the elements of $\gamma(A \perp x)$ as the *best* elements of $A \perp x$, whenever the latter is nonempty. Note that in this definition, $\leq$ is required to be fixed for all choices of x; otherwise all partial meet contraction functions would be trivially relational. Note also that the definition does not require any special properties of $\leq$ apart from being a relation; if there is a transitive relation $\leq$ such that for all $x \notin \mathrm{Cn}(\varnothing)$ the marking off identity holds, then γ is said to be *transitively relational* over A. Finally, we say that a partial meet contraction function $\dot{-}$ is relational (transitively relational) over A iff it is determined by some selection function that is so. "Some", because a single partial meet contraction function may, in the infinite case, be determined by two distinct selection functions. In the finite case, however, this cannot happen, as we shall show in Observation 4.6.

Relationality is linked with supplementary postulate ($\dot{-}7$), and transitive relationality even more closely linked with the conjunction of ($\dot{-}7$) and ($\dot{-}8$). Indeed, we shall show, in the first group of results of this section, that a partial meet contraction function $\dot{-}$ is transitively relational iff ($\dot{-}7$) and ($\dot{-}8$) are both satisfied. In the later part of this section we shall describe the rather more complex relationship between relationality and ($\dot{-}7$) considered alone. It will be useful to consider various further conditions, and two that are of immediate assistance are:

($\gamma 7$) $\gamma(A \perp x \,\&\, y) \subseteq \gamma(A \perp x) \cup \gamma(A \perp y)$ for all x and y.
($\gamma 8$) $\gamma(A \perp x) \subseteq \gamma(A \perp x \,\&\, y)$ whenever $A \perp x \cap \gamma(A \perp x \,\&\, y) \neq \varnothing$.

As with ($\dot{-}8$), it is easy to show that when A is a theory and γ is a selection function over A, then ($\gamma 8$) can equivalently be formulated as

$$\gamma\,(A \perp x) \subseteq \gamma\,(A \perp x \,\&\, y) \;\text{ whenever } A \perp x \cap \gamma(A \perp y) \neq \varnothing.$$

The following lemma will also be needed throughout the section.

Lemma 4.1 *Let A be any theory and $x, y \in A$. Then $A \perp (x \,\&\, y) = A \perp x \cup A \perp y$.*

Proof We apply Lemma 2.4. When $B \in A \perp (x \,\&\, y)$, then $B \nvdash x \,\&\, y$ so $B \nvdash x$ or $B \nvdash y$, so by 2.4 either $B \in A \perp x$ or $B \in A \perp y$. Conversely, if $B \in A \perp x$ or $B \in A \perp y$, then $B \nvdash x \,\&\, y$ so, by 2.4 again, $B \in A \perp (x \,\&\, y)$. □

Observation 4.2 *Let A be a theory and $\dot{-}$ a partial meet contraction function over A determined by a selection function γ. If γ satisfies the condition (γ7), then $\dot{-}$ satisfies ($\dot{-}$7), and if it satisfies (γ8), then $\dot{-}$ satisfies ($\dot{-}$8).*

Proof Suppose (γ7) holds. Then we have:

$$\begin{aligned}(A\dot{-}x) \cap (A\dot{-}y) &= \bigcap \gamma\,(A \perp x) \cap \bigcap \gamma\,(A \perp y) \text{ since } \gamma \text{ determines } \dot{-}\\ &= \bigcap (\gamma\ (A \perp x) \cup \gamma\,(A \perp y)) \text{ by general set theory}\\ &\subseteq \bigcap \gamma\,(A \perp (x\&y)) \text{ using condition } (\gamma 7)\\ &= A\dot{-}(x\&y)\,.\end{aligned}$$

Suppose now that (γ8) holds, and suppose $x \notin A \dot{-} (x \,\&\, \gamma)$; that is, $x \notin \bigcap\gamma(A \perp x \,\&\, y)$. We need to show that $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x$. In the case $x \notin A$ we have $A \dot{-} (x \,\&\, y) = A = A \dot{-} x$. So suppose $x \in A$. Since $x \notin \bigcap\gamma(A \perp x \,\&\, y)$ there is a $B \in \gamma(A \perp x \,\&\, y)$ with $B \nvdash x$, so, by Lemma 2.4, $B \in A \perp x$ and thus $B \in A \perp x \cap \gamma(A \perp x \,\&\, y)$. Applying ($\gamma$8) we have $\gamma(A \perp x) \subseteq \gamma(A \perp x \,\&\, y)$, so $A \dot{-} (x \,\&\, y) = \bigcap\gamma(A \perp x \,\&\, y) \subseteq \bigcap\gamma(A \perp x) = A \dot{-} x$ as desired. □

Observation 4.3 *Let A be any theory and γ a selection function for A. If γ is relational over A then γ satisfies the condition (γ7), and if γ is transitively relational over A, then γ satisfies the condition (γ8).*

Proof In the cases that $x \in \mathrm{Cn}(\varnothing)$, $y \in \mathrm{Cn}(\varnothing)$, $x \notin A$ and $y \notin A$, both (γ7) and (γ8) hold trivially, so we may suppose that $x \notin \mathrm{Cn}(\varnothing)$, $y \notin \mathrm{Cn}(\varnothing)$, $x \notin A$ and $y \in A$.

Suppose γ is relational over A, and suppose $B \in \gamma(A \perp x \,\&\, y)$. Now $\gamma(A \perp x \,\&\, y) \subseteq A \perp x \,\&\, y = A \perp x \cup A \perp y$, so $B \in A \perp x$ or $B \in A \perp \gamma$; consider the former case, as the latter is similar. Let $B' \in A \perp x$. Then $B' \in A \perp x \cup A \perp y = A \perp x \,\&\, y$, and so $B' \leq B$ since $B \in \gamma\ (A \perp x \,\&\, y)$ and γ is relational over A; and thus, by relationality again, $B' \in (A \perp x) \subseteq \gamma(A \perp x) \cup \gamma(A \perp y)$ as desired.

Suppose now that γ is transitively relational over A, and suppose $A \perp x \cap \gamma(A \perp x \,\&\, y) \neq \varnothing$. Suppose for reductio ad absurdum that there is a $B \in \gamma\ (A \perp x)$ with $B \notin y\ (A \perp x \,\&\, y)$. Since $B \in \gamma\ (A \perp x) \subseteq A \perp x \subseteq A \perp x \,\&\, y$ by Lemma 4.1, whilst $B \notin \gamma\ (A \perp x \,\&\, y)$, we have by relationality that there is a $B' \in A \perp x \,\&\, y$ with $B' \leq B$. Now by the hypothesis $A \perp x \cap y\ (A \perp x \,\&\, y) \neq \varnothing$, there is a $B'' \in A \perp x$ with $B'' \in \gamma(A \perp x \,\&\, y)$. Hence by relationality $B' \leq B''$ and also $B'' \leq B$. Transitivity gives us $B' \leq B$ and thus a contradiction. □

When A is a theory and γ is a selection function for A, we define γ^*, the *completion* of γ, by putting $\gamma^*(A \perp x) = \{B \in A \perp x\colon \bigcap\gamma(A \perp x) \subseteq B\}$ for all $x \notin \mathrm{Cn}(\varnothing)$, and $\gamma^*(A \perp x) = \gamma(A \perp x) = \{A\}$ in the limiting case that $x \in \mathrm{Cn}(\varnothing)$.

It is easily verified that γ^* is also a selection function for A, and determines the same partial meet contraction function as γ does. Moreover, we clearly have $\gamma\ (A \perp x) \subseteq \gamma^*(A \perp x) = \gamma^{**}(A \perp x)$ for all x. This notion is useful in the formulation of the following statement:

Observation 4.4 *Let A be any theory, and $\dot{-}$ a partial meet contraction function over A, determined by a selection function γ. If $\dot{-}$ satisfies the conditions* ($\dot{-}7$) *and* ($\dot{-}8$) *then γ^* is transitively relational over A.*

Proof Define the relation $\leq$ over 2^A as follows: for all $B, B' \in 2^A$, $B' \leq B$ iff either $B' = B = A$, or the following three all hold:

(i) $B' \in A \perp x$ for some $x \in A$.
(ii) $B \in A \perp x$ and $A \dot{-} x \subseteq B$ for some $x \in A$.
(iii) For all x, if $B', B \in A \perp x$ and $A \dot{-} x \subseteq B'$, then $A \dot{-} x \subseteq B$.

We need to show that the relation is transitive, and that it satisfies the marking off identity $\gamma^*(A \perp x) = \{B \in A \perp x: B' \leq B \text{ for all } B' \in A \perp x\}$ for all $x \notin \mathrm{Cn}(\varnothing)$.

For the identity, suppose first that $B \in \gamma^*(A \perp x) \subseteq A \perp x$ since $x \notin \mathrm{Cn}(\varnothing)$. Let $B' \in A \perp x$; we need to show that $B' < B$. If $x \notin A$ then $B' = B = A$ so $B' \leq B$. Suppose that $x \in A$. Then clearly conditions (i) and (ii) are satisfied. Let y be any proposition, and suppose $B', B \in A \perp y$ and $A \dot{-} y \subseteq B'$; we need to show that $A \dot{-} y \subseteq B$. Now by covering, which we have seen to follow from ($\dot{-}8$), either $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x$ or $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} y$. And in the latter case $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} y \subseteq B' \in A \perp x$ so $x \notin A \dot{-} (x \,\&\, \gamma)$; so by ($\dot{-}8$) again $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x$. Thus in either case $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x$. Now suppose for reductio ad absurdum that there is a $w \in A \dot{-} y$ with $w \notin B$. Then $y \vee w \in A \dot{-} y$ and so since $y \vdash y \vee w$ we have by ($\dot{-}7$) using Observation 3.3 that $y \vee w \in A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x = \bigcap \gamma^*(A \perp x) \subseteq B$; so $y \vee w \in B$. But also since $B \in A \perp y$ and $w \notin B$ and $w \in A$, we have $B \cup \{w\} \vdash y$, so $\neg w \vee y \in B$. Putting these together gives us $(y \vee w) \,\&\, (y \vee \neg w) \in B$, so $y \in B$, contradicting $B \in A \perp y$.

For the converse, suppose $B \notin \gamma^*(A \perp x)$ and $B \in A \perp x$; we need to find a $B' \in A \perp x$ with $B' \not\leq B$. Clearly the supposition implies that $x \in A$, so $B \neq A$. Since $B \in A \perp x$, the latter is nonempty, so $\gamma^*(A \perp x)$ is nonempty; let B' be one of its elements. Noting that $B', B \in A \perp x$, $B' \in \gamma^*(A \perp x)$, but $B \notin y^*(A \perp x)$, we see that condition (iii) fails, so that $B' \not\leq B$, as desired.

Finally, we check out transitivity. Suppose $B'' \leq B'$ and $B' \leq B$; we want to show that $B'' \leq B$. In the case that $B = A$ then clearly since $B' \leq B$ we have $B' = B = A$, and thus since $B'' \leq B'$ we have $B'' = B' = A$, so $B'' = B = A$ and $B'' \leq B$. Suppose for the principal case that $B \neq A$. Then since $B' \leq B$, clearly $B' \neq A$. Since $B' \leq B$ we have $B \in A \perp w$ and $A \dot{-} w \subseteq B$ for some $w \in A$, so (ii) is satisfied. Since $B'' \leq B'$ we have $B'' \in A \perp w$ for some $w \in A$, so (i) is satisfied. It remains to verify (iii). Suppose $B'', B \in A \perp y$ and $A \dot{-} y \subseteq B''$; we need to show that $A \dot{-} y \subseteq B$. First, note that since $B \neq A$ by the condition of the case, we have $y \in A$. Also, since $B'' \leq B'$ and $B' \neq A$, there is an $x \in A$ with $B' \in A \perp x$ and $A \dot{-} x \subseteq B'$. Since $x, y \in A$ we have by Lemma 4.1 that $A \perp x \,\&\, y = A \perp x \cup A \perp y$, so $B'', B', B \in A \perp x \,\&\, y$. Now by covering, either $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} y$ or $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x$. The former case gives us $A \dot{-} (x \,\&\, y) \subseteq B''$, so since $B'' \leq B'$ and $B' \neq A$ We have $A \dot{-} (x \,\&\, y) \subseteq B'$,

so again since $B' \leq B$ and $B \neq A$ we have $A \dot{-} (x \,\&\, y) \subseteq B$. Likewise, the latter case gives us $A \dot{-} (x \,\&\, y) \subseteq B'$, so since $B' \leq B$ and $B \neq A$ we have $A \dot{-} (x \,\&\, y) \subseteq B$. Thus in either case, $A \dot{-} (x \,\&\, y) \subseteq B$. Now let $w \in A \dot{-} y$ we need to show that $w \in B$. Since $w \in A \dot{-} y$ we have $y \vee w \in A \dot{-}$ y; so by ($\dot{-}7$) and Observation 3.3, since $y \vee w \in \mathrm{Cn}(y)$, we have $y \vee w \in A \dot{-} (x \,\&\, y) \subseteq B$. Hence $B \cup \{\neg\, y\} \vdash w$. But since $B \in A \perp y$ and $w \in A$, we also have $B \cup \{y\} \vdash w$, so $B \vdash w$ and thus $w \in B$ as desired. □

Corollary 4.5 *Let A be any theory, and* $\dot{-}$ *a partial meet contraction function over A determined by a selection function* γ. *Then* $\dot{-}$ *is transitively relational over A iff* $\dot{-}$ *satisfies both* ($\dot{-}7$) *and* ($\dot{-}8$).

Proof If $\dot{-}$ satisfies ($\dot{-}7$) and ($\dot{-}8$) then, by 4.4, γ^* is transitively relational, so since γ^* determines $\dot{-}$, the latter is transitively relational. Conversely, if $\dot{-}$ is transitively relational, then γ' is transitively relational for some γ' that determines $\dot{-}$; so, by 4.3, γ' satisfies ($\gamma 7$) and ($\gamma 8$); so, by 4.2, $\dot{-}$ satisfies ($\dot{-}7$) and ($\dot{-}8$). □

This result is the promised representation theorem for the collection of "basic" plus "supplementary" postulates. Since this collection of postulates can be independently motivated (cf. Gärdenfors (1978)), there is strong reason to focus on transitively relational partial meet contraction functions as an ideal representation of the intuitive process of contraction.

Note that Observation 4.4 and its corollary give us a sufficient condition for the transitive relationality of γ^*, and thus of $\dot{-}$, rather than of γ itself. The question thus arises: when can we get the latter? We shall show that in the finite case the passage from γ to $\dot{-}$ is injective, so that $\gamma = \gamma^*$, where γ is any selection function that determines $\dot{-}$. By the *finite case*, we mean the case where A is finite modulo Cn; that is, where the equivalence relation defined by $\mathrm{Cn}(x) = \mathrm{Cn}(y)$ partitions A into finitely many cells.

Observation 4.6 *Let A be any theory finite modulo* Cn, *and let* γ *and* γ' *be selection functions for A. For every proposition x, if* $\gamma\ (A \perp x) \neq \gamma'\ (A \perp x)$, *then* $\bigcap \gamma(A \perp x) \neq \cap \gamma'(A \perp x)$.

Sketch of Proof Suppose $B \in \gamma(A \perp x)$, but $B \notin \gamma'(A \perp x)$. Then clearly $x \in A$ and $x \notin \mathrm{Cn}(\varnothing)$. Since A is finite (we identify A with its quotient structure), so is B; put b to be the conjunction of its elements. Then it is easy to check that $b \in B$ but $b \notin B'$ for all $B' \notin \gamma'\ (A \perp x)$. Put $c = \neg b \vee x$: then it is easy to check that $c \notin B \supseteq \bigcap \gamma(A \perp x)$, but $c \in B'$ for all $B' \in \gamma'(A \perp x)$; that is, $c \in \bigcap \gamma'(A \perp x)$. □

Corollary 4.7 *Let A be any theory finite modulo* Cn, *and* $\dot{-}$ *a partial meet contraction function over A determined by a selection function* γ. *If* $\dot{-}$ *satisfies conditions* ($\dot{-}7$) *and* ($\dot{-}8$), *then y is transitively relational over A.*

Proof Immediate from 4.4 and 4.6. □

We turn now to the question of the relation of condition ($\dot{-}7$), considered alone, to relationality; and here the situation is rather more complex and less satisfying. Now we have from Observation 4.2 that when $\dot{-}$ is determined by γ, then if γ satisfies

($\gamma 7$), then $\dot{-}$ satisfies ($\dot{-}7$), and it is not difficult to show, by an argument similar to that of 4.6, that:

Observation 4.8 *If A is a theory finite modulo* Cn, *and $\dot{-}$ a partial meet contraction function over A determined by a selection function γ, then $\dot{-}$ satisfies* ($\dot{-}7$) *iff γ satisfies* ($\gamma 7$). *Also, $\dot{-}$ satisfies* ($\dot{-}8$) *iff γ satisfies* ($\gamma 8$).

But on the other hand, even in the finite case, ($\gamma 7$) does *not* imply the relationality of γ or of $\dot{-}$:

Observation 4.9 *There is a theory A, finite modulo* Cn, *with a partial meet contraction function $\dot{-}$ over A, determined by a selection function γ, such that $\dot{-}$ satisfies* ($\gamma 7$), *but $\dot{-}$ is not relational over A.*

Sketch of Proof Take the sixteen-element Boolean algebra, take an atom a_0 of this algebra, and put A to be the principal filter determined by a_0. This will be an eight-element structure, lattice-isomorphic to the Boolean algebra of eight elements. We take Cn in the natural way, putting $\mathrm{Cn}(X) = \{x\colon \wedge X \leq x\}$. We label the eight elements of A as $a_0, \ldots, a_7$, where a_0 is already defined, a_1, a_2, a_3 are the three atoms of A (not of the entire Boolean algebra), a_4, a_5, a_6 are the three dual atoms of A, and a_7 is the greatest element of A (i.e. the unit of the Boolean algebra). For each $i \leq 7$, we write $!a_i$ for $\{a_j \in A\colon a_i \leq a_j\}$. We define γ by putting $\gamma(A \perp a_7) = \gamma(A \perp \mathrm{Cn}(\varnothing)) = \{A\} = !a_0$ as required in this limiting case, $\gamma(A \perp a_0) = A \perp a_j$ for all j with $1 \leq j < 7$, and $\gamma(A \perp a_0) = \{!a_1\}$. Then it is easy to verify that for all $a_i \notin \mathrm{Cn}(\varnothing)$, $\gamma(A \perp a_i)$ is a nonempty subset of $A \perp a_i$, so γ is a selection function for A. By considering cases we easily verify ($\gamma 7$) (and thus also by 4.2 ($\dot{-}7$)); and by considering the role of $!a_2$ it is easy to verify that γ (and hence by 4.6, $\dot{-}$ itself) is not relational over A. □

The question thus arises whether there is a condition on $\dot{-}$ or on γ that is equivalent to the relationality of $\dot{-}$ or of γ respectively. We do not know of any such condition for $\dot{-}$, but there is one for γ, of an infinitistic nature. It is convenient, in this connection, to consider a descending series of conditions, as follows:

(γ7:∞) $A \perp x \cap \bigcap_{i\in I}\{\gamma(A \perp y_i)\} \subseteq \gamma(A \perp x)$, whenever $A \perp x \subseteq \bigcup_{i\in I}\{A \perp y_j\}$.
(γ7:N) $A \perp x \cap \gamma(A \perp y_1) \cap \cdots \cap \gamma(A \perp y_n) \subseteq \gamma(A \perp x)$, whenever $A \perp x \subseteq A \perp y_1 \cup \cdots \cup A \perp y_n$, for all $n \geq 1$.
(γ7:2) $A \perp x \cap \gamma(A \perp y_1) \cap \gamma(A \perp y_2) \subseteq \gamma(A \perp x)$, whenever $A \perp x \subseteq A \perp y_1 \cup A \perp y_2$.
(γ7:1) $A \perp x \cap \gamma(A \perp y) \subseteq \gamma(A \perp x)$, whenever $A \perp x \subseteq A \perp y$.

Observation 4.10 *Let A be any theory and γ a selection function over A. Then γ is relational over A iff* (γ7: ∞) *is satisfied. Moreover, we have* (γ7:∞) $\rightarrow$ (γ7:N) $\leftrightarrow$ (γ7:2) $\rightarrow$ (γ7:1) $\leftrightarrow$ (γ7). *On the other hand,* (γ7:1) *does not imply* (γ7:2), *even in the finite case; although in the finite case,* (γ7:N) *is equivalent to* (γ7: ∞).

Sketch of Proof Writing (γR) for "γ is relational over A", we show first that (γR) $\rightarrow$ (γ7:∞). Suppose (γR), and suppose $A \perp x \subseteq \bigcup_{i\in I}\{A \perp y_i\}$. Suppose $B \in A \perp x \cap \bigcap_{i\in I}\{\gamma(A \perp y_i)\}$. We need to show that $B \in \gamma(A \perp x)$. Since $B \in \gamma(A \perp y_i)$

for all $i \in I$, we have by relationality that $B' \leq B$ for all $B' \in A \perp y_i$, for all $i \in I$; so, by the supposition, $B' \leq B$ for all $B' \in A \perp x$. Hence, since $B \in A \perp x$, so that also $x \notin \mathrm{Cn}(\varnothing)$, we have by relationality that $B \in \gamma(A \perp x)$. To show the converse ($\gamma 7$: ∞) $\rightarrow$ (γR), suppose ($\gamma 7$: ∞) holds, and define $\leq$ over 2^A by putting $B' \leq B$ iff there is an x with $B \in \gamma(A \perp x)$ and $B' \in A \perp x$; we need to verify the marking off identity. The left to right inclusion of the marking off identity is immediate. For the right to left, suppose $B \in A \perp x$ and for all $B' \in A \perp x$, $B' \leq B$. Then by the definition of $\leq$, for all $B_i \in \{B_i\}_{i \in I} = A \perp x$ there is a y_i with $B \in \gamma(A \perp y_i)$ and $B_i \in A \perp y_i$. Since $B_i \in A \perp y_i$, for all $B_i \in A \perp x$, we have $A \perp x \subseteq \bigcup_{i \in I}\{A \perp y_i\}$, so we may apply ($\gamma 7$: ∞). But clearly $B \in A \perp x \cap \bigcap_{i \in I}\{\gamma(A \perp y_i)\}$. Hence by ($\gamma 7$:$\infty$) we have $B \in \gamma(A \perp x)$, as desired.

The implications ($\gamma 7$:∞) $\rightarrow$ ($\gamma 7$:N) $\rightarrow$ ($\gamma 7$:2) $\rightarrow$ ($\gamma 7$:1) are trivial, as is the equivalence of ($\gamma 7$:∞) to ($\gamma 7$:N) in the finite case. To show that ($\gamma 7$:2) implies the more general ($\gamma 7$:N), it suffices to show that for all $n \geq 2$, ($\gamma 7$:n) $\rightarrow$ ($\gamma 7$:$n + 1$): this can be done using the fact that when $y_n, y_{n+1} \in A$, $A \perp y_n \cup A \perp y_{n+1} = A \perp (y_n \,\&\, y_{n+1})$ by Lemma 4.1.

To show that ($\gamma 7$:1) $\rightarrow$ ($\gamma 7$), recall from 4.1 that when $x, y \in A$, then $A \perp x \,\&\, y = A \perp x \cup A \perp y$; so $A \perp x \subseteq A \perp x \,\&\, y$, and so, by ($\gamma 7$:1), $(A \perp x) \cap \gamma\ (A \perp x \,\&\, y) \subseteq \gamma(A \perp x)$. Similarly $(A \perp y) \cap \gamma(A \perp x \,\&\, y) \subseteq \gamma(A \perp y)$. Forming unions on left and right, distributing on the left, and applying 4.1 gives us $\gamma(A \perp x \,\&\, y) \subseteq \gamma(A \perp x) \cup \gamma(A \perp y)$ as desired.

To show conversely that ($\gamma 7$) $\rightarrow$ ($\gamma 7$:1), suppose ($\gamma 7$) is satisfied, suppose $A \perp x \subseteq A \perp y$ and consider the principal case that $x, y \in A$. Then using compactness we have $y \vdash x$, so $\mathrm{Cn}(y) = \mathrm{Cn}(x \,\&\, (\neg x \vee y))$, so by ($\gamma 7$)

$$\gamma(A \perp y) \subseteq \gamma(A \perp x) \cup \gamma(A \perp \neg x \vee y)\,,$$

so $A \perp x \cap \gamma(A \perp y) \subseteq \gamma(A \perp x) \cup \gamma(A \perp \neg x \vee y)$. The verification is then completed by showing that $A \perp x$ is disjoint from $\gamma\ (A \perp \neg x \vee y)$.

Finally, to show that ($\gamma 7$:1) does not imply ($\gamma 7$:2), even in the finite case, consider the same example as in the proof of Observation 4.9. We know from that proof that this example satisfies ($\gamma 7$) and thus also ($\gamma 7$:1), but that γ is not relational over A, so by earlier parts of this proof, ($\gamma 7$:∞) fails, so by finiteness ($\gamma 7$:N) fails, so ($\gamma 7$:2) fails. Alternatively, a direct counterinstance to ($\gamma 7$:2) in this example can be obtained by putting $x = a_0$, $y_1 = a_1$, and $y_2 = a_2$. □

5 Remarks on Connectivity

It is natural to ask what the consequences are of imposing connectivity as well as transitivity on the relation that determines a selection function. Perhaps surprisingly, it turns out that in the infinite case it adds very little, and in the finite case nothing at all. This is the subject of the present section.

Let A be a set of propositions and γ a selection function for A. We say that γ is *connectively relational* over A iff there is a relation that is connected over 2^A such that for all $x \notin \mathrm{Cn}(\varnothing)$, the marking off identity of Sect. 4 holds. And a partial meet contraction function is called connectively relational iff it is determined by some selection function that is so.

We note as a preliminary that it suffices to require connectivity over the much smaller set $U_A = \bigcup_{x \in A}\{A \perp x\}$. For suppose that $\leq$ is connected over U_A. Put $\leq_0$ to be the restriction of $\leq$ to U_A; then $\leq_0$ will still be connected over U_A. Then put $\leq_1$ to be $\leq_0 \cup ((2^A - U_A) \times 2^A)$. Clearly $\leq_1$ will be connected over 2^A. Moreover, if $\leq$ satisfies the marking off identity, so does $\leq_1$.

Indeed, when $\leq$ is transitive, it suffices to require connectivity on the even smaller set $U_\gamma = \bigcup\{\gamma(A \perp x){:}x \in A, x \notin \mathrm{Cn}(\varnothing)\}$. For here likewise we can define $\leq_0$ as the restriction of $\leq$ to U_γ, and then define $\leq_1$, to be $\leq_0 \cup ((2^A - U_\gamma) \times 2^A)$. Then clearly $\leq_1$ is connected over 2^A, and is transitive if $\leq$ is transitive; and we can easily check, using the transitivity of $\leq$, that if $\leq$ satisfies the marking off identity for γ, so does $\leq_1$.

Observation 5.1 *Let A be any theory and $\dot{-}$ a partial meet contraction function over A. Then $\dot{-}$ is transitively relational iff it is transitively and connectively relational.*

Proof Suppose that $\dot{-}$ is determined by the transitively relational selection function γ. Then by 4.2 and 4.3, $\dot{-}$ satisfies the conditions ($\dot{-}7$) and ($\dot{-}8$), so the conditions of Observation 4.4 hold and the relation $\leq$ defined in its proof is transitive and satisfies the marking off identity for γ^*. By the above remarks, to show that $\dot{-}$ is transitively and connectively relational it suffices to show that $\leq$ is connected over the set U_{γ^*}.

Let $B', B \in U_{\gamma^*}$ and suppose $B' \not\leq B$. Since $B', B \in U_{\gamma^*}$, conditions (i) and (ii) of the definition of $\leq$ in the proof of 4.4 are satisfied for both B' and B. Hence since $B' \leq B$ we have by (iii) that there is an x with $B', B \in A \perp x$, $A \dot{-} x \subseteq B'$ and $A \dot{-} x \not\subseteq B$. But since $A \dot{-} x \subseteq B' \in A \perp x$ we have by the definition of γ^* that $B' \in \gamma^*(A \perp x)$, so by the marking off identity for γ^*, $\leq$ as verified in the proof of 4.4, since $B \in A \perp x$ we have $B \leq B'$ as desired. □

In the case that A is finite modulo Cn, this result can be both broadened and sharpened: Broadened to apply to relationality in general rather than only to transitive relationality, and sharpened to guarantee connectivity over U_A of any given relation under which the selection function γ is relational, rather than merely connectivity, as above, of a specially constructed relation under which the closure γ^* of γ is relational.

Observation 5.2 *Let A be a theory finite modulo* Cn, *and let $\dot{-}$ be a partial meet contraction function over A, determined by a selection function γ. Suppose that γ is relational, with the relation $\leq$ satisfying the marking off identity. Then $\leq$ is connected over U_A.*

Proof Let $B', B \in U_A = \bigcup_{x \in A}\{A \perp x\}$. Since A is finite modulo Cn, there are $b', b \in A$ with $A \perp b' = \{B'\}$ and $A \perp b = \{B\}$—for example, put b to be the disjunction of

all (up to equivalence modulo Cn) the elements $a \in A$ such that $B \nvdash a$. Now $A \perp b' \,\&\, b = A \perp b' \cup A \perp b = \{B', B\}$ by Lemma 4.1, and so since γ is a selection function, $\gamma(A \perp b' \,\&\, b)$ is a nonempty subset of $\{B', B\}$, which implies that either B' or B is in $\gamma(A \perp b' \,\&\, b)$. In the former case we have $B \leq B'$, and in the latter case we have the converse. □

Corollary 5.3 *Let A be a theory finite modulo* Cn, *and let $\dot{-}$ be a partial meet contraction function over A. Then $\dot{-}$ is relational iff it is connectively relational.*

Proof Immediate from 5.2. □

6 Maxichoice Contraction Functions and Factoring Conditions on $A \dot{-} (x \,\&\, y)$

The first topic of this section will be a brief investigation of the consequences of the following rather strong *fullness* condition:

($\dot{-}$F) If $y \in A$ and $y \notin A \dot{-} x$, then $\neg y \vee x \in A \dot{-} x$, for any theory A.

From the results in Gärdenfors (1982), it follows that if $\dot{-}$ is a partial meet contraction function, then this condition (there called (−6)) is equivalent with the following condition (called (21) in Gärdenfors (1982)) on partial meet revision functions:

($\dot{+}$ F) If $y \in A$ and $y \notin A \dot{+} x$, then $\neg y \in A \dot{+} x$, for any theory A.

The strength of the condition ($\dot{-}$F) is shown by the following simple representation theorem:

Observation 6.1 *Let $\dot{-}$ be any partial meet contraction function over a theory A. Then $\dot{-}$ satisfies ($\dot{-}$F) iff $\dot{-}$ is a maxichoice contraction function.*

Proof Suppose $\dot{-}$ satisfies ($\dot{-}$F). Suppose B, $B' \in \gamma(A \perp x)$ and assume for contradiction that $B \neq B'$. There is then some $\nu \in B'$ such that $y \notin B$. Hence $y \notin A \dot{-} x$ and since y $\in A$ it follows from ($\dot{-}$F) that $\neg\, y \vee x \in A \dot{-} x$. Hence $\neg\, y \vee x \in B'$, but since $y \in B'$ it follows that $x \in B'$, which contradicts the assumption that $B' \in A \perp x$. We conclude that $B = B'$ and hence that $\dot{-}$ is a maxichoice contraction function.

For the converse, suppose that $\dot{-}$ is a maxichoice contraction function and suppose that $y \in A$ and $y \in A \dot{-} x$. Since $A \dot{-} x = B$ for some $B \in A \perp x$, it follows that $y \notin B$. So by the definition of $A \perp x$, $x \in \mathrm{Cn}(B \cup \{\mathrm{y}\})$. By the properties of the consequence operation we conclude that $\neg y \vee x \in B = A \dot{-} x$, and thus ($\dot{-}$F) is satisfied.

In addition to this representation theorem for maxichoice contraction functions, we can also prove another one based on the following *primeness* condition.

(∸Q) For all $y, z \in A$ and for all x, if $y \vee z \in A \dot{-} x$, then either $y \in A \dot{-} x$ or $z \in A \dot{-} x$.

Observation 6.2 *Let $\dot{-}$ be any partial meet contraction function over a theory A. Then $\dot{-}$ satisfies (∸Q) iff $\dot{-}$ is a maxichoice contraction function.*

Proof Suppose first that $\dot{-}$ is a maxichoice function and suppose $y, z \in A$, $y \notin A \dot{-} x$ and $z \notin A \dot{-} x$. Then by maximality, $A \dot{-} x \cup \{y\} \vdash x$ and $A \dot{-} - x \cup \{z\} \vdash$, so $A \dot{-} x \cup \{y \vee z\} \vdash x$. But since say $y \in A$ and $y \notin A \dot{-} x$, we have $x \notin \mathrm{Cn}(\emptyset)$, so $x \notin A \dot{-} x$. Thus $y \vee z \notin A \dot{-} x$, which shows that (∸Q) is satisfied.

For the converse, suppose that (∸Q) is satisfied. By Observation 6.1, it suffices to derive (∸F). Suppose $y \in A$ and $y \notin A \dot{-} x$. We need to show that $\neg y \vee x \in A \dot{-} x$. Now $(y \vee \neg y) \vee x \notin \mathrm{Cn}(\emptyset)$, and so $(y \vee \neg y) \vee x = y \vee (\neg y \vee x) \in A \dot{-} x$. Also by hypothesis $y \in A$, and since $y \notin A \dot{-} x$ we have $x \in A$, so $\neg y \vee x \in A$. We can now apply the primeness condition (∸Q) and get either $y \in A \dot{-} x$ or $\neg y \vee x \in A \dot{-} x$. By hypothesis, the former fails, so the latter holds and (∸F) is verified.

With the aid of these results we shall now look at three "factoring" conditions on the contraction of a conjunction from a theory A. They are

Decomposition (∸D). For all x and y, $A \dot{-} (x \,\&\, y) = A \dot{-} x$ or $A \dot{-} (x \,\&\, y) = A \dot{-} y$.
Intersection (∸I). For all x and y in A, $A \dot{-} (x \,\&\, y) = A \dot{-} x \cap A \dot{-} y$.
Ventilation (∸V). For all x and y, $A \dot{-} (x \,\&\, y) = A \dot{-} x$ or $A \dot{-} (x \,\&\, y) = A \dot{-} y$ or $A \dot{-} (x \,\&\, y) = A \dot{-} x \cap A \dot{-} y$.

These bear some analogy to the very processes of maxichoice, full meet, and partial meet contraction respectively, and the analogy is even more apparent if we express the factoring conditions in their equivalent n-ary forms:

$$A \dot{-} (x_1 \& \cdots \& x_n) = A \dot{-} x_i \quad \text{for some } i \leq n;$$
$$A \dot{-} (x_1 \& \cdots \& x_n) = \bigcap_{i \leq n} \{A \dot{-} x_i\} \quad \text{whenever } x_1, \ldots, x_n \in A;$$
$$A \dot{-} (x_1 \& \cdots \& x_n) = \bigcap_{i \in I} \{A \dot{-} x_i\} \quad \text{for some } I, \text{ where } \emptyset \neq I \subseteq \{1, \ldots n\}.$$

This analogy of formulation corresponds indeed to quite close relationships between the three kinds of contraction process, on the one hand, and the three kinds of factorization on the other. We shall state the essential relationships first, to give a clear overall picture, and group the proofs together afterwards.

First, the relationship between maxichoice contraction and decomposition. In (Alchourron and Makinson 1982b) it was shown that if A is a theory and $\dot{-}$ is a maxichoice contraction function over A, then decomposition is equivalent to each of (∸7) and (∸8). In the more general context of partial meet contraction functions these equivalences between the conditions break down, and it is decomposition (∸D) that emerges as the strongest among them:

Observation 6.3 *Let A be a theory and $\dot{-}$ a partial meet contraction function over A. Then the following conditions are equivalent:*

(a) $\dot{-}$ *satisfies* (∸D).

(b) $\dot{-}$ *is a maxichoice contraction function and satisfies at least one of* ($\dot{-}$7) *and* ($\dot{-}$8).
(c) $\dot{-}$ *is a maxichoice contraction function and satisfies both of* ($\dot{-}$7) *and* ($\dot{-}$8).
(d) $\dot{-}$ *satisfies* ($\dot{-}$WD).
(e) $\dot{-}$ *is a maxichoice contraction function and satisfies* ($\dot{-}$C).

Here ($\dot{-}$WD) is the *weak decomposition* condition: for all x and y, $A \dot{-} x \subseteq A \dot{-} x \,\&\, y$ or $A \dot{-} y \subseteq A \dot{-} x \,\&\, y$.

The relationship of full meet contraction to the intersection condition ($\dot{-}$I) is even more direct. This is essentially because a full meet contraction function, as defined at the beginning of Sect. 2, is always transitively relational, and so always satisfies ($\dot{-}$7) and ($\dot{-}$8). For since $\gamma(A \perp x) = A \perp x$ for all $x \notin \mathrm{Cn}(\emptyset)$, γ is determined via the marking off identity by the total relation over 2^A or over $\bigcup x\{A \perp x\}$.

Observation 6.4 *Let A be a theory and $\dot{-}$ a partial meet contraction function over A. Then the following conditions are equivalent:*

(a) $\dot{-}$ *satisfies* ($\dot{-}$I).
(b) $\dot{-}$ *satisfies* ($\dot{-}$M).
(c) $\dot{-}$ *is a full meet contraction function.*

Here ($\dot{-}$M) is the *monotony condition:* for all $x \in A$, if $x \vdash y$ then $A \dot{-} x \subseteq A \dot{-} y$. This result gives us a representation theorem for full meet contraction. Note, as a point of detail, that whereas decomposition and ventilation are formulated for arbitrary propositions x and y, the intersection and monotony conditions are formulated under the restriction that x and y (respectively, x) are in A. For if $x \notin A$, then $x \,\&\, y \notin A$, so $A \dot{-} (x \,\&\, y) = A$ whilst $A \dot{-} x \cap A \dot{-} y = A \cap A \dot{-} y = A \dot{-} y \neq A$ if $y \in A$ and $y \notin \mathrm{Cn}(\emptyset)$.

Of the three factoring conditions, ventilation ($\dot{-}$V) is clearly the most "general" and the weakest. But it is still strong enough to imply the "supplementary postulates" ($\dot{-}$7) and ($\dot{-}$8):

Observation 6.5 *Let A be a theory and $\dot{-}$ a partial meet contraction function over A. Then $\dot{-}$ satisfies* ($\dot{-}$V) *iff $\dot{-}$ satisfies both* ($\dot{-}$7) *and* ($\dot{-}$8).

Proof of Observation 6.3 We know by the chain of equivalences in §8 of Alchourron and Makinson (1982b) that if $\dot{-}$ is a maxichoice contraction function then the conditions ($\dot{-}$7), ($\dot{-}$8) and ($\dot{-}$D) are mutually equivalent. This already shows the equivalence of (b) and (c), and also shows that they imply (a). (d) is a trivial consequence of (a). To prove the equivalence of (a)–(d) it remains to show that (d) implies (b).

Suppose that $\dot{-}$ satisfies ($\dot{-}$WD). Clearly it then satisfies ($\dot{-}$7), so we need only verify that $\dot{-}$ is a maxichoice function, for which it suffices by Observation 6.1 to verify ($\dot{-}$F); that is, that whenever $y \in A$ and $y \notin A \dot{-} x$ then $\neg\, y \vee x \in \mathrm{A} \dot{-} x$. Suppose for reductio ad absurdum that $y \in A$, $y \notin A \dot{-} x$ and $\neg\, y \vee x \notin \mathrm{A} \dot{-} x$. Note that this implies that $x \in A$. Now $\mathrm{Cn}(x) = \mathrm{Cn}((x \vee y) \,\&\, (x \vee \neg\, y))$, so by ($\dot{-}$WD) we have $A \dot{-} (x \vee y) \subseteq A \dot{-} x$ or $A \dot{-} (x \vee \neg\, y) \subseteq A \dot{-} x$. In the former case, $\neg\, y \vee x \in A \dot{-} (x \vee y)$. But by recovery $A \dot{-} (x \vee y) \cup \{x \vee y\} \vdash x$, so $A \dot{-} (x \vee y) \cup \{y\} \vdash x$, so

$\neg$ y $\vee$ $x \in A \dot{-} (x \vee y)$, giving a contradiction. And in the latter case, $y \notin A \dot{-} (x \vee \neg y)$, whereas by recovery $A \dot{-} (x \vee \neg y) \cup \{x \vee \neg y\} \vdash y$, so $A \dot{-} (x \vee \neg y) \cup \{\neg y\} \vdash y$, so $y \in A \dot{-} (x \vee \neg y)$, again giving a contradiction.

Finally, it must be shown that (e) is equivalent with (a)–(d). First note that it follows immediately from Observation 3.4 that (c) entails (e). To complete the proof we show that (e) entails (b). In the light of Observation 6.1 it suffices to show that ($\dot{-}$F) and ($\dot{-}$C) together entail ($\dot{-}$8). To do this assume that $x \notin A \dot{-} x \,\&\, y$. We want to show that $A \dot{-} x \,\&\, y \subseteq A \dot{-} x$. In the case when $x \notin A$, this holds trivially; so suppose that $x \in A$. It then follows from ($\dot{-}$F) that $\neg x \vee (x \,\&\, y) \in A \dot{-} x \,\&\, y$, so $\neg x \vee y \in A \dot{-} (x \,\&\, y) = A \dot{-} (\neg x \vee y) \,\&\, x$. By ($\dot{-}$C), $A \dot{-} x \,\&\, y = A \dot{-} (\neg x \vee y) \,\&\, x \subseteq A \dot{-} x \vee y$ or $A \dot{-} x \,\&\, y = A \dot{-} (\neg x \vee y) \,\&\, x \subseteq A \dot{-} x$. Since the second case is the desired inclusion, it will suffice to show that the first case implies the second. Suppose $A \dot{-} x \,\&\, y \subseteq A \dot{-} \neg x \vee y$. Then, since $\neg x \vee y \in A \dot{-} x \,\&\, y$, we have $\neg x \vee y \in A \dot{-} \neg x \vee y$, so by ($\dot{-}$4) $\neg x \vee y \in \mathrm{Cn}(\emptyset)$. But this means that $\mathrm{Cn}(x \,\&\, y) = \mathrm{Cn}(x)$, so $A \dot{-} x \,\&\, y = A \dot{-} x$ by ($\dot{-}$5), and we are done. □

The last part of this proof shows that for maxichoice contraction functions the converse of Observation 3.4 also holds.

Proof of Observation 6.4 Suppose first that $\dot{-}$ is a full meet contraction function. We show that ($\dot{-}$I) is satisfied. If $x \in \mathrm{Cn}(\emptyset)$ or $y \in \mathrm{Cn}(\emptyset)$ then the desired equation holds trivially. Suppose that $x, y \notin \mathrm{Cn}(\emptyset)$, and suppose that $x, y \in A$. Then we may apply Observation 4.1 to get

$$A \dot{-} (x \& y) = \bigcap \{A \perp x \& y\} = \bigcap \{A \perp x \cup A \perp y\} = \bigcap \{A \perp x\} \cap \bigcap \{A \perp y\}$$
$$= A \dot{-} x \cap A \dot{-} y,$$

so that $\dot{-}$ satisfies the intersection condition.

Trivially, intersection implies monotony. Suppose that $\dot{-}$ satisfies monotony; to prove (c) we need to show that $A \dot{-} x = A \sim x$, for which it clearly suffices to show that $A \dot{-} x \subseteq A \sim x$ in the light of Observation 2.7. In the case $x \notin A$ this holds trivially. In the case $x \in A$ we have by Observation 2.1 of (Alchourron and Makinson 1982b) that $A \sim x = A \cap \mathrm{Cn}(\neg x)$, so we need only show $A \dot{-} x \subseteq \mathrm{Cn}(\neg x)$. Suppose $y \in A \dot{-} x$. Then by ($\dot{-}$M), since $x \in A$ and $x \vdash x \vee y$, we have $y \in A \dot{-} (x \vee y)$, so $x \vee y \in A \dot{-} (x \vee y)$; so, by the postulate ($\dot{-}$4), $x \vee y \in \mathrm{Cn}(\emptyset)$, so that $y \in \mathrm{Cn}(\neg x)$ as desired. □

Proof of Observation 6.5 For the left to right implication, suppose $\dot{-}$ satisfies ($\dot{-}$V). Then ($\dot{-}$7) holds immediately. For ($\dot{-}$8), let x and y be propositions and suppose $x \notin A \dot{-} (x \,\&\, y)$; we need to show that $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x$. In the case that $x \notin A$ this holds trivially, so we suppose $x \in A$. Now $\mathrm{Cn}(x \,\&\, y) = \mathrm{Cn}(x \,\&\, (\neg x \vee y))$, so by ($\dot{-}$V) $A \dot{-} (x \,\&\, y)$ is identical with one of $A \dot{-} x$, $A \dot{-} (\neg x \vee y;)$ or $(A \dot{-} x) \cap (A \dot{-} (\neg x \vee y))$. In the first and last cases we have the desired inclusion, so we need only show that the middle case is impossible. Now by recovery, $A \dot{-} (\neg x \vee y) \cup \{\neg x \vee y\} \vdash x$, so $A \dot{-} (\neg x \vee y) \cup \{\neg x\} \vdash x$, so $x \in A \dot{-} (\neg x \vee y)$. But by hypothesis, $x \notin A \dot{-} (x \,\&\, y)$, so $A \dot{-} (x \,\&\, y) \neq A \dot{-} (\neg x \vee y)$, as desired.

The converse can be proven via the representation theorem (Observation 4.4), but it can also be given a direct verification as follows. Suppose that $\dot{-}$ satisfies ($\dot{-}$7) and ($\dot{-}$8), and suppose that $A \dot{-} (x \,\&\, y) \neq A \dot{-} x$ and $A \dot{-} (x \,\&\, y) \neq A \dot{-} y$; we want to show that $A \dot{-} (x \,\&\, y) = A \dot{-} x \cap A \dot{-} y$. By ($\dot{-}$7) it suffices to show that $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} x \cap A \dot{-} y$, so it suffices to show that $A \dot{-} (x \,\&\, y) \subseteq \dot{-} x$ and $A \dot{-} (x \,\&\, y) \subseteq A \dot{-} y$. By ($\dot{-}$C), which we know by 3.4 to be an immediate consequence of ($\dot{-}$8), we have at least one of these inclusions. So it remains to show that under our hypotheses either inclusion implies the other. We prove one; the other is similar.

Suppose for reductio ad absurdum that $A \dot{-} (x \,\&\, y) \subseteq \dot{-} x$ but $A \dot{-} (x \,\&\, y) \not\subseteq A \dot{-} y$. Since by hypothesis $A \dot{-} (x \,\&\, y) \neq A \dot{-} x$, we have $A \dot{-} x \not\subseteq A \dot{-} (x \,\&\, y)$, so there is an $a \in A \dot{-} x$ with $a \notin A \dot{-} (x \,\&\, y)$. Since $A \dot{-} (x \,\&\, y) \not\subseteq A \dot{-} y$, we have by ($\dot{-}$8) that $y \in A \dot{-} (x \,\&\, y)$. Hence since $a \notin A \dot{-} (x \,\&\, y)$ we have $\neg\, y \vee a \notin A \dot{-} (x \,\&\, y)$. Hence by ($\dot{-}$7), $\neg\, y \vee a \notin A \dot{-} x$ or $\neg\, y \vee a \notin A \dot{-} y$. But since $a \in A \dot{-} x$ the former alternative is impossible. And the second alternative is also impossible, since by recovery $A \dot{-} y \cup \{y\} \vdash a$, so that $\neg\, y \vee a \in A \dot{-} y$. □

7 A Diagram for the Implications

To end the paper, we summarize the "implication results" of Sects. 4 and 6 in a diagram. The conditions are as named in previous pages with in addition ($\dot{-}$R) and ($\dot{-}$TR), meaning that $\dot{-}$ is relational, respectively transitively relational, over A, and (γR) and (γTR), meaning that γ is. ($\dot{-}$C) is the covering condition of Observation 3.4; (γC) is its analogue $\gamma(A \perp x) \subseteq \gamma(A \perp x \,\&\, y)$ or $\gamma(A \perp y) \subseteq \gamma(A \perp x \,\&\, y)$. ($\dot{-}$ P) is the partial antitony condition of 3.3; and (γP) is its obvious analogue $\gamma(A \perp x \,\&\, y) \cap A \perp x \subseteq \gamma(A \perp x)$. Conditions are understood to be formulated for an arbitrary theory A, selection function γ for A, and partial meet contraction function $\dot{-}$ over A determined by γ. Arrows are of course for implications, and conditions grouped into the same box are mutually equivalent in the finite case. Conversely, conditions in separate boxes are known to be nonequivalent, even for the finite case. The diagram should be read as a map of an ordering, but *not* as a lattice: a "$\vee$" alignment does not necessarily mean that the bottom condition is equivalent to the conjunction of the other two. In some cases, it is—for example ($\dot{-}$TR) = ($\dot{-}$7) & ($\dot{-}$8) = ($\dot{-}$V), as proven in Corollary 4.5 and Observation 6.5; and again ($\dot{-}$D) = ($\dot{-}$F) & ($\dot{-}$8), as shown in Observation 6.3. But ($\dot{-}$7) & ($\dot{-}$C) is known *not* to be equivalent to ($\dot{-}$TR), and (γR) & ($\dot{-}$TR) may perhaps not be equivalent to (γTR). Finally, implications and nonimplications that follow from others by transitivity have not been written into the diagram, but are left as understood. Implications concerning connectivity from Sect. 5 have been omitted from the diagram, to avoid overcluttering.

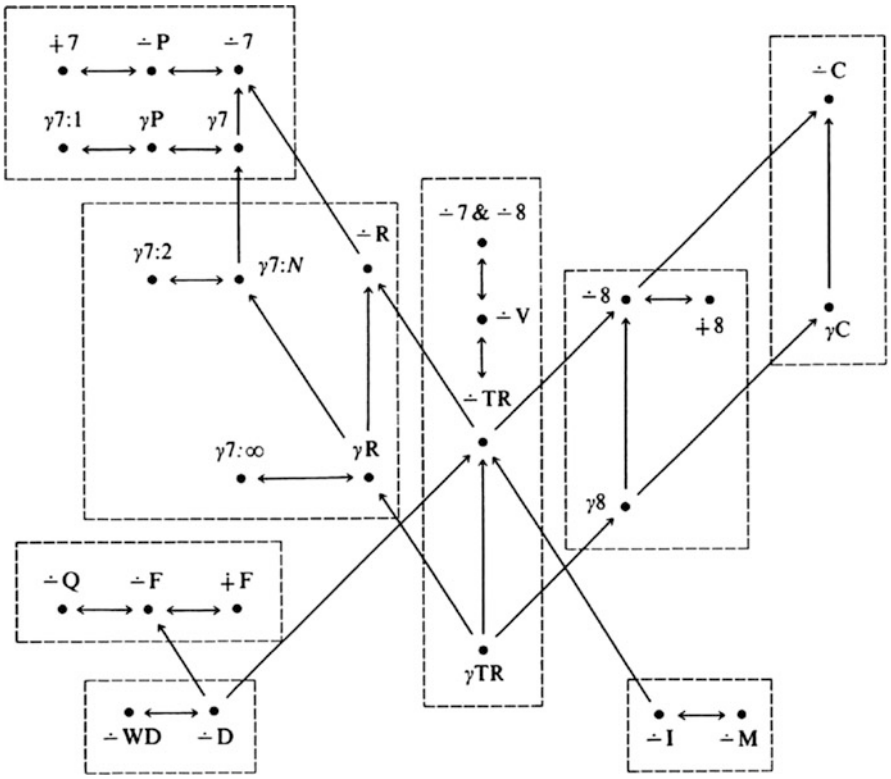
All the general implications (arrows) have been proven in the text, or are immediate. The finite case equivalences issue from the injection result of Observation 4.6, and several were noted in Observation 4.10. Of the finite case non-equivalences, a first example serving to separate (γ7) from ($\dot{-}$R) was given in Observation 4.9, from

which it follows immediately that ($\gamma 7$) does not in the finite case imply ($\dot{-}$TR). The other nonequivalences need other examples, which we briefly sketch.

For the second example, take A to be the eight-element theory of Observation 4.9, but define γ as follows: In the limiting case of a_7, we put $\gamma(A \perp a_7) = \{!a_0\}$ as required by the fact that $a_7 \in \mathrm{Cn}(\varnothing)$; put $\gamma(A \perp a_j) = A \perp a_j$ for all j with $2 \leq j < 7$; put $\gamma(A \perp a_1) = \{!a_3\}$; and put $\gamma(A \perp a_0) = \{!a_1, !a_3\}$. Then it can be verified that the partial meet contraction function $\dot{-}$ determined by γ satisfies ($\dot{-}$C), and so by finiteness also (γC), but not ($\dot{-}8$) and so a fortiori not ($\dot{-}$TR).

For the third example, take A as before, and put $\gamma(A \perp a_7) = \{!a_0\}$ as always; put $\gamma(A \perp a_1) = \{!a_2\}$; and put $\gamma(A \perp a_i) = A \perp a_i$ for all other a_i It is then easy to check that this example satisfies ($\dot{-}8$) but not ($\dot{-}7$), and so a fortiori not ($\dot{-}$R) and not ($\dot{-}$TR).

For the fourth and last example, take A as before, and put $\leq$ to be the least reflexive relation over 2^A such that $!a_1 \leq !a_2$, $!a_2 \leq !a_3$, $!a_3 \leq !a_2$ and $!a_3 \leq !a_1$. Define γ from $\leq$ via the marking off identity, and put $A \dot{-} x = \bigcap \gamma(A \perp x)$. Then it is easy to check that γ is a selection function for A, so that ($\dot{-}$R) holds. But ($\dot{-}$C) fails; in particular when $x = a_1$ and $y = a_2$ we can easily verify that $A \dot{-} (x \,\&\, y) \not\subseteq A \dot{-} x$ and $A \dot{-} (x \,\&\, y) \not\subseteq A \dot{-} y$. Hence, a fortiori, ($\dot{-}8$) and ($\dot{-}$TR) also fail.



8 Added in Proof

The authors have obtained two refinements: the arrow $(\dot{-}\mathrm{D}) \rightarrow (\dot{-}\mathrm{TR})$ of the diagram on page 528 can be strengthened to $(\dot{-}\mathrm{D}) \rightarrow (\gamma\mathrm{TR})$; the implication $(\gamma 7{:}\infty) \rightarrow (\gamma 7{:}N)$ of Observation 4.10 can be strengthened to an equivalence. The former refinement is easily verified using the fact that any *maxichoice* contraction function over a theory is determined by a *unique* selection function over that theory. The latter refinement can be established by persistent use of the compactness of Cn.

Observation 4.10 so refined implies that for a theory A and selection function γ over A, γ is relational over A iff $(\gamma 7{:}2)$ holds. This raises an interesting *open question*, a positive answer to which would give a representation theorem for relational partial meet contraction, complementing Corollary 4.5: Can condition $(\gamma 7{:}2)$ be expressed as a condition on the contraction operation $\dot{-}$ determined by γ?

We note that a rather different approach to contraction has been developed by Alchourrón and Makinson in *On the logic of theory change: safe contraction*, to appear in ***Studia Logica*****,** vol. 44 (1985), the issue dedicated to Alfred Tarski; the relationship between the two approaches is studied by the same authors in *Maps between some different kinds of contraction function: the finite case*, also to appear in ***Studia Logica*****,** vol. 44 (1985).

References

Alchourron, C. E., & Makinson, D. (1982a). *Hierarchies of regulations and their logic, new studies in deontic logic* (pp. 125–148), R. Hilpinen (Ed.), Dordrecht: Reidel.

Alchourron, C. E., & Makinson, D. (1982b). On the logic of theory change: Contraction functions and their associated revision functions. *Theoria, 48*, 14–37.

Gärdenfors, P. (1978). Conditionals and changes of belief The logic and epistemology of scientific change. In I. Niiniluoto & R. Tuomela (Eds.), *Acta Philosophica Fennica* (Vol. 30, pp. 381–404).

Gärdenfors, P. (1982). *Rules for rational changes of belief 320311: Philosophical essays dedicated to Lennart Åqvist on his fiftieth birthday* (pp. 88–101), (T. Pauli, Ed., Philosophical studies no. 34). Uppsala: Department of Philosophy, University of Uppsala.

Gärdenfors, P. (1984). Epistemic importance and minimal changes of belief. *Australasian Journal of Philosophy, 62*, 136–157.

Makinson, D. (1985). How to give it up: A survey of some formal aspects of the logic of theory change. *Synthese, 62*, 347–363.

Chapter 14
Theory Contraction and Base Contraction Unified

Sven Ove Hansson

General Introduction

The AGM (Alchourrón-Gärdenfors-Makinson) model of belief change has acquired the status of a standard model. In that model, a belief state is represented by a set of sentences that is closed under logical consequence, the *belief set* or theory. (Alchourrón et al. 1985); (Gärdenfors 1988) Among the major rivals are models in which a belief state is represented by a *belief base*, a set that is not (except in a limiting case) closed under consequence. (Fuhrmann 1991); (Hansson 1989; 1991; 1992; 1993); (Nebel 1992) Obviously, the logical closure of a belief base is a belief set, and each non-trivial belief set can be represented by several different belief bases. It has been argued that different belief bases for one and the same belief set represent different ways of holding the same beliefs. Roughly, the elements of the belief base represent "basic", or independently grounded beliefs, in contrast to the "merely derived" beliefs that form the rest of the belief set. (Fuhrmann 1991, pp. 183–184); (Hansson 1989; 1992)

As has been accurately pointed out by Fuhrmann (1991), operations on a belief base generate operations on the corresponding belief set. In particular, if a belief base for a belief set **K** is contracted by a sentence α, then the logical closure of the contracted belief base is a belief set from which α has been contracted.

The purpose of this paper is to characterize the operators of contraction on a belief set **K** that can be obtained by assigning to it (1) a belief base, and (2) an operator of partial meet contraction for that belief base. The section

Sven Ove Hansson 1992-07-08.

S.O. Hansson (✉)
Division of Philosophy, KTH, Brinellvägen 32, 100 44 Stockholm, Sweden
e-mail: soh@kth.se

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_14

"Partial meet contraction" contains introductory material. In the section "The new postulates", the postulates that will be used for the characterizations are introduced, and in the section "Axiomatic characterizations" axiomatic characterizations of various types of base-generated theory contractions are given. The section "Proofs" provides proofs of the results reported in the section "Axiomatic characterizations".

Partial Meet Contraction

We will assume a language L that is closed under truth-functional operations and a consequence operator Cn for L. Cn satisfies the standard conditions for a consequence operator, namely inclusion ($A \subseteq Cn(A)$), monotony (if $A \subseteq B$, then $Cn(A) \subseteq Cn(B)$), and iteration ($Cn(A) = Cn(Cn(A))$). Furthermore, it satisfies the properties of supraclassicality (if α follows by classical truth-functional logic from A, then $\alpha \in Cn(A)$), deduction (if $\beta \in Cn(A \cup \{\alpha\})$), then $(\alpha \rightarrow \beta) \in Cn(A)$) and compactness (if $\alpha \in Cn(A)$, then $\alpha \in Cn(A')$ for some finite set $A' \subseteq A$). $A \vdash \alpha$ will be used as an alternative notation for $\alpha \in Cn(A)$.

A *belief set* is a subset $\mathbf{K}$ of the language such that $\mathbf{K} = Cn(\mathbf{K})$. An operator of *contraction* for $\mathbf{K}$ is an operator $\div$ from $\mathcal{P}(L) \times L$ to $\mathcal{P}(L)$ such that for all sentences α, $\mathbf{K} \div \alpha \subseteq \mathbf{K}$, and if $\alpha \notin Cn(\varnothing)$, then $\alpha \notin Cn(\mathbf{K} \div \alpha)$. A particularly interesting type of contraction is *partial meet contraction*, which was introduced by Alchourrón, Gärdenfors, and Makinson (1985). (Gärdenfors 1984) It is defined by the identity:

$$\mathbf{K} \div \alpha = \bigcap \gamma(\mathbf{K} \perp \alpha),$$

where $\mathbf{K} \perp \alpha$ denotes the set of inclusion-maximal subsets of $\mathbf{K}$ that do not have α as a logical consequence. γ is a *selection function*, such that $\gamma(\mathbf{K} \perp \alpha)$ is a non-empty subset of $\mathbf{K} \perp \alpha$, unless the latter is empty, in which case $\gamma(\mathbf{K} \perp \alpha) = \{\mathbf{K}\}$.

Let γ be a selection function for $\mathbf{K}$. Then the *completion* of γ is the function γ^* such that $\gamma^*(\mathbf{K} \perp \alpha) = \{X \in \mathbf{K} \perp \alpha | \bigcap \gamma(\mathbf{K} \perp \alpha) \subseteq X\}$, unless $\mathbf{K} \perp \alpha$ is empty, in which case $\gamma^*(\mathbf{K} \perp \alpha) = \{\mathbf{K}\}$. γ^* is a selection function for $\mathbf{K}$, and gives rise to the same operator of partial meet contraction as γ. (Alchourrón et al. 1985, p. 519) It is in some contexts convenient to make the technical assumption that a selection function is *completed*, i.e. identical to its own completion.

Full meet contraction is the limiting case when $\gamma(\mathbf{K} \perp \alpha) = \mathbf{K} \perp \alpha$ for all non-empty $\mathbf{K} \perp \alpha$. In the other limiting case, when $\gamma(\mathbf{K} \perp \alpha)$ is a singleton for all α, $\div$ is an operator of *maxichoice* contraction. An operator of partial meet contraction is *relational* if and only if it is based on a selection function γ for which there is a relation $\underline{\ll}$ such that for all non-empty $\mathbf{K} \perp \alpha$ we have

$$\gamma(\mathbf{K} \perp \alpha) = \left\{\mathbf{K}' \in \mathbf{K} \perp \alpha \;\middle|\; \mathbf{K}'' \underline{\ll} \mathbf{K}' \text{ for all } \mathbf{K}'' \in \mathbf{K} \perp \alpha\right\}$$

If this condition holds for a transitive relation $\underline{\ll}$, then the operator is *transitively relational*.

Partial meet contraction derives much of its attractiveness from a representation theorem by Alchourrón et al. (1985): An operation ÷ on a logically closed set **K** is a partial meet contraction if and only if it satisfies the following six postulates, the basic *Gärdenfors postulates*:

(G÷1) $\mathbf{K} \div \alpha$ is a theory if **K** is a theory (*closure*)
(G÷2) $\mathbf{K} \div \alpha \subseteq \mathbf{K}$ (*inclusion*)
(G÷3) If $\alpha \notin Cn(\mathbf{K})$ then $\mathbf{K} \div \alpha = \mathbf{K}$ (*vacuity*)
(G÷4) If $\alpha \notin Cn(\varnothing)$ then $\alpha \notin Cn(\mathbf{K} \div \alpha)$ (*success*)
(G÷5) If $Cn(\alpha) = Cn(\beta)$ then $\mathbf{K} \div \alpha = \mathbf{K} \div \beta$ (*preservation*)
(G÷6) $\mathbf{K} \subseteq Cn((\mathbf{K} \div \alpha) \cup \{\alpha\})$ whenever **K** is a theory (*recovery*).

Furthermore, an operator of partial meet contraction on a logically closed set is transitively relational if and only if it also satisfies:

(G÷7) $(\mathbf{K} \div \alpha) \cap (\mathbf{K} \div \beta) \subseteq \mathbf{K} \div (\alpha \& \beta)$ (*intersection*)
(G÷8) If $\alpha \notin \mathbf{K} \div (\alpha \& \beta)$ then $\mathbf{K} \div (\alpha \& \beta) \subseteq \mathbf{K} \div \alpha$ (*conjunction*).

A *belief base* for **K** is a set $\boldsymbol{B}$ such that $\mathbf{K} = Cn(\boldsymbol{B})$. Partial meet contraction for belief bases is defined in the same way as for belief sets, i.e., by the identity $\boldsymbol{B} \div \alpha = \bigcap \gamma(\boldsymbol{B} \perp \alpha)$. Full meet, maxichoice, relational, and transitively relational contraction is also defined in the same way as for belief sets. Furthermore, an operator of partial meet contraction is *transitively, maximizingly relational* (*TMR*) if and only if it is relational by a transitive relation $\underline{\ll}$ such that, for its strict counterpart «, if $A \subset B$, then $A \ll B$.[1]

Contractions on belief bases may be studied in their own right.[2] In this paper, however, they will be treated as means for contractions on belief sets.

Definition An operation ÷ on a belief set **K** is *generated by* a partial meet base contraction if and only if there is a belief base $\boldsymbol{B}$ for **K** and an operator $\sim_\gamma$ of partial meet contraction for $\boldsymbol{B}$ such that $\mathbf{K} \div \alpha = Cn\left(\boldsymbol{B} \sim_\gamma \alpha\right)$ for all $\alpha \in L$.

We will see directly that if an operation on a belief set is generated by some partial meet base contraction, then it satisfies the first five of the basic Gärdenfors postulates, (G÷1)–(G÷5), but it does not, in general, satisfy recovery (G÷6).[3]

[1] The maximizing property may be interpreted as saying that all elements of the belief base have positive epistemic value. This property might at first hand seem superfluous. If $K' \subset K''$, then $K' \in K \perp \alpha$ and $K'' \in K \perp \alpha$ cannot both be true, so that K' and K'' cannot be candidates between which the selection function has to choose. However, when more than one contraction is taken into account, the property need not be superfluous. If $K_1 \subset K_2$, and K_3 is neither a subset nor a superset of either K_1 or K_2, then $\gamma(\{K_1, K_3\}) = \{K_1\}$ and $\gamma(\{K_2, K_3\}) = \{K_3\}$ may both hold for a transitively relational selection function γ, but not if it is in addition required to be maximizing.

[2] See (Hansson 1993) for some results, including an axiomatic characterization of the partial meet contractions on a belief base.

[3] In Makinson's (1987) terminology, an operation that satisfies (G÷1)-(G÷5) but not necessarily (G÷6) is a "withdrawal". However, I will use "contraction" in the wide sense indicated above, thus including Makinson's withdrawals.

Recovery is the most controversial of the six postulates, and has been seriously questioned by several authors. See (Hansson 1991); (Makinson 1987); (Niederée 1991).

The New Postulates

It has often been remarked that the only realistic belief sets are those that have a finite representation. (Gärdenfors and Makinson 1988) We are going to assume that both the original belief set and the belief sets that are obtained through contraction have finite representations (belief bases); thus for every α, there is some finite set A such that $\mathbf{K} \div \alpha = Cn(A)$. Furthermore, we will assume that although there may be infinitely many sentences by which the belief set can be contracted, there is only a finite number of belief sets that can be obtained through contraction, i.e. $\{\mathbf{K}' | (\exists\alpha)(\mathbf{K}' = \mathbf{K} \div \alpha)\}$ is finite. These two finiteness properties can be combined into the following:

There is a finite set A such that for every α, there is some $A' \subseteq A$ such that $\mathbf{K} \div \alpha = Cn(A')$. (*finitude*)

In the presence of vacuity (G÷3), *finitude* implies that there is some finite set A such that $\mathbf{K} = Cn(A)$.

If $\div$ is a contraction operator for $\mathbf{K}$, and α is not a logical theorem, then $\mathbf{K} \div \alpha$ does not contain α. However, to contract $\mathbf{K}$ by α is not the only way to exclude α from the belief set. Typically, there are several β distinct from α such that $\alpha \notin \mathbf{K} \div \beta$. This must be the case if α logically implies β, and it can be true in other cases as well. A contraction $\mathbf{K} \div \beta$ such that $\alpha \notin \mathbf{K} \div \beta$ will be called an α-*removal*.

Two different beliefs may have exactly the same justification(s). As an example, I believe that either Paris or Oslo is the capital of France (α). I also believe that either Paris or Stockholm is the capital of France (β). Both these beliefs are entirely based on my belief that Paris is the capital of France. Therefore, a contraction by some sentence δ removes α if and only if it removes β (namely if and only if it removes the common justification of these two beliefs). There is no contraction by which I can retract α without retracting β or vice versa. It is not unreasonable to require that if two beliefs in a belief set stand or fall together in this way, then their contractions are identical. In other words, *if all* α*-removals are* β*-removals and vice versa, then* $\mathbf{K} \div \alpha = \mathbf{K} \div \beta$. In the formal language:

If $\mathbf{K} \div \delta \vdash \alpha$ iff $\mathbf{K} \div \delta \vdash \beta$ for all δ, then $\mathbf{K} \div \alpha = \mathbf{K} \div \beta$. (*symmetry*)

An α-removal $\mathbf{K} \div \beta$ will be called a *preservative* α*-removal* if and only if $\mathbf{K} \div \alpha \subseteq \mathbf{K} \div \beta$, and a *strictly preservative* α*-removal* if and only if $\mathbf{K} \div \alpha \subset \mathbf{K} \div \beta$. A strictly preservative α-removal is an operation that removes α, and does this in a more economical way than what is done by the contraction by α.

Often, a belief can be removed more economically if more specified information is obtained. As an example, I believe that Albert Schweitzer was a German Missionary (α). Let α_1 denote that he was a German and α_2 that he was a Missionary,

so that $\alpha \equiv \alpha_1 \& \alpha_2$. If I have to contract my belief set by α, then the contracted belief set will contain neither α_1 nor α_2. Admittedly it would be logically sufficient to withdraw one of them. However, they are both equally entrenched, so that I do not know which to choose in preference over the other. Therefore, both will have to go. On the other hand, if I have to contract my belief set by α_1, then I have no reason to let go of α_2.[4] To contract by α_1 is, given the structure of my belief state, a more specified way to remove α. Thus we may expect that $\mathbf{K} \div \alpha \subset \mathbf{K} \div \alpha_1$, so that $\mathbf{K} \div \alpha_1$ is a strictly preservative α-removal.

Let δ denote that Albert Schweitzer was a Swede, and let us consider the contraction of $\mathbf{K}$ by $\alpha_1 \vee \delta$, "Albert Schweitzer was a German or a Swede." Since I believe in $\alpha_1 \vee \delta$ only as a consequence of my belief in α_1, I can only retract $\alpha_1 \vee \delta$ by retracting α_1. Therefore, $\mathbf{K} \div (\alpha_1 \vee \delta)$ is not a proper superset of $\mathbf{K} \div \alpha_1$, i.e., it is not a more conservative α-withdrawal than $\mathbf{K} \div \alpha_1$. Indeed, the way my beliefs are structured, α_1 cannot be further subdivided in the way that α was subdivided into α_1 and α_2. There is no part of α_1 that stands on its own and can be retracted from $\mathbf{K}$ without the rest of α_1 being lost as well. In this sense, no α-removal can be more conservative than $\mathbf{K} \div \alpha_1$.

More generally, $\mathbf{K} \div \beta$ is a *maximally preservative* α*-removal* if and only if it is a preservative α-removal and there is no α-removal $\mathbf{K} \div \delta$ such that $\mathbf{K} \div \beta \subset \mathbf{K} \div \delta$. Intuitively, to perform a maximally preservative α-removal is to make the belief set not imply α, making use of information that is sufficiently specified to allow one to remove a part of α so small that no smaller part of it can be removed alone.

Contraction should be conservative in the sense that every element of $\mathbf{K}$ is retained in $\mathbf{K} \div \alpha$ unless there is some good reason to exclude it. As was noted in (Hansson 1991), (G÷1)–(G÷5) do not ensure this property, since they are satisfied by the operation $\div$ such that if $\alpha \notin Cn(\mathbf{K})$, then $\mathbf{K} \div \alpha = \mathbf{K}$, and otherwise $\mathbf{K} \div \alpha = \mathbf{K} \cap Cn(\varnothing)$. The same applies if *symmetry* is added to the list of postulates. We need some further postulate to prevent elements of $\mathbf{K}$ from being lost for no reason in operations of contraction.

One way to achieve this is to require that what is lost in the contraction by α must be incompatible with some reasonable way to remove α. In other words, if a unit of belief is lost in contraction by α, then there should be at least one preservative α-removal to which the lost unit cannot be added without α being implied. However, it would be too far-reaching to require this to hold for units of belief that cannot stand on their own (such as $\alpha_1 \vee \delta$ in our Albert Schweizer example). Such a unit of belief can be lost merely due to loss of the larger, self-sustained unit(s) of which it is a part. Thus, our principle of conservativity should refer only to units of belief that *can* stand on their own.

In order to formulate a conservativity principle, we therefore need a formal criterion that excludes non-self-sustained units of belief. In our example, there can be no sentence β such that $\mathbf{K} \div \beta$ consists exactly of $\alpha_1 \vee \delta$ and its logical consequences, since $\alpha_1 \vee \delta$ cannot stand on its own without α_1. Admittedly, this is a

[4] It is here assumed that $\alpha_1 \equiv \alpha_2$ is less entrenched than α_1 and α_2.

weak criterion, and it can be strengthened in various ways.[5] However, it turns out to be sufficient for our present purposes, and it will therefore be used in our postulate of conservativity:

If $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$, then there is some δ such that $\mathbf{K} \div \alpha \subseteq \mathbf{K} \div \delta \not\vdash \alpha$ and $(\mathbf{K} \div \beta) \cup (\mathbf{K} \div \delta) \vdash \alpha$. (*conservativity*)

An obvious way to strengthen *conservativity* is to require that if a unit of belief is lost in contraction by α, then α will be implied if it is added to $\mathbf{K} \div \alpha$ (and not merely if it is added to some preservative α-removal):

If $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$ then $\mathbf{K} \div \alpha \not\vdash \alpha$ and $(\mathbf{K} \div \beta) \cup (\mathbf{K} \div \alpha) \vdash \alpha$. (*strong conservativity*)

Strong conservativity is much less plausible than *conservativity*. This can be seen from our Albert Schweitzer example. In that example, it may reasonably be assumed that $\alpha_1 \notin \mathbf{K} \div (\alpha_1 \& \alpha_2)$, $\alpha_2 \notin \mathbf{K} \div (\alpha_1 \& \alpha_2)$, $\alpha_1 \in \mathbf{K} \div \alpha_2$, $\alpha_2 \in \mathbf{K} \div \alpha_1$, $\mathbf{K} \div (\alpha_1 \& \alpha_2) \subseteq \mathbf{K} \div \alpha_1$, and $\mathbf{K} \div (\alpha_1 \& \alpha_2) \subseteq \mathbf{K} \div \alpha_2$. However, this is incompatible with *strong conservativity*. Since $\mathbf{K} \div \alpha_1 \not\subseteq \mathbf{K} \div (\alpha_1 \& \alpha_2)$, this postulate requires that $\mathbf{K} \div \alpha_1 \cup \mathbf{K} \div (\alpha_1 \& \alpha_2) \vdash (\alpha_1 \& \alpha_2)$, contrary to our assumptions for this case. More generally, *strong conservativity* is implausible since it precludes the removal of two or more sentences (in this case α_1 and α_2), when it would have been logically sufficient to remove only one of them. Such epistemic behaviour is rational enough when the beliefs in question are equally entrenched, or have equal epistemic utility.

The concepts of epistemic entrenchment and epistemic utility refer to extra-logical reasons that one may have for preferring one way to remove a sentence α rather than another. It is conceivable for an epistemic agent to make no use of such extra-logical information. Such an agent is indecisive in the sense of not being able to make a choice among different ways to remove a belief, if these are on an equal footing from a logical point of view. Formally, this is close to a reversal of *conservativity*: If a (self-sustained) unit of belief conflicts with some way to remove α from $\mathbf{K}$, then it is not a part of $\mathbf{K} \div \alpha$:

If there is some δ such that $\mathbf{K} \div \delta \not\vdash \alpha$ and $(\mathbf{K} \div \beta) \cup (\mathbf{K} \div \delta) \vdash \alpha$, then $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$. (*indecisiveness*)

Non-logical considerations play an important role in actual human epistemic behaviour. Arguably, it would in many cases be irrational not to let them do so. Therefore, *indecisiveness* is not a plausible general property of rational belief change.

[5] Two such strengthened versions should be mentioned: (1) $\div$ can be extended to contraction by sets (multiple contraction), such that if $A \cap Cn(\varnothing) = \varnothing$, then $\mathbf{K} \div A$ is a logically closed subset of $\mathbf{K}$ that does not imply any element of A. It a unit of belief cannot stand on its own, then it should not be equal to $\mathbf{K} \div A$ for any set A. (2) Iterated contraction can be used for the same purpose: If a unit of belief cannot stand on its own, then it should not be equal to $\mathbf{K} \div \beta_1 \div \beta_2 \ldots \div \beta_n$ for any series $\beta_1 \ldots \beta_n$ of sentences.

The next postulate will again be concerned with maximally preservative removals, and thus with the smallest units of belief that can be removed from a belief set. In an orderly and coherent belief state, one would expect the identities and the relations of entrenchment of these units to be constant, i.e. they should be the same independently of what sentence we are contracting by.

As an illustration, let us extend the Albert Schweitzer example. Let α_1 denote that Schweitzer was a German, α_2 that he was a missionary and α_3 that he was a physician. Let us assume that $\mathbf{K} \div \alpha_1$ is a maximally preservative $\alpha_1 \,\&\, \alpha_2 \,\&\, \alpha_3$-removal, i.e. a maximally economical way to remove $\alpha_1 \,\&\, \alpha_2 \,\&\, \alpha_3$ from the belief set. Since $\mathbf{K} \div \alpha_1$ is also an $\alpha_1 \,\&\, \alpha_2$-removal, and since the $\alpha_1 \,\&\, \alpha_2$-removals are a subset of the $\alpha_1 \,\&\, \alpha_2 \,\&\, \alpha_3$-removals, it should also be maximally preservative among these. Furthermore, if $\mathbf{K} \div \alpha_2$ is equally economical as $\mathbf{K} \div \alpha_1$ in the context of removing $\alpha_1 \,\&\, \alpha_2$ (i.e., if it is also a maximaly preservative $\alpha_1 \,\&\, \alpha_2$-removal), then it should also be equally economical as $\mathbf{K} \div \alpha_1$ in the context of removing $\alpha_1 \,\&\, \alpha_2 \,\&\, \alpha_3$ (i.e., it should also be a maximally preservative $\alpha_1 \,\&\, \alpha_2 \,\&\, \alpha_3$-removal). In general:

If $\vdash \alpha \rightarrow \beta$ and the set of β-removals that are also maximally preservative α-removals is non-empty, then it coincides with the set of maximally preservative β-removals. (*regularity*)

It may be convenient to divide *regularity* into two parts[6]:

If $\vdash \alpha \rightarrow \beta$ and some maximally preservative α-removal is also a β-removal, then all maximally preservative β-removals are maximally preservative α-removals. (*regularity 1*)

If $\vdash \alpha \rightarrow \beta$ and $\mathbf{K} \div \delta$ is both a β-removal and a maximally preservative α-removal, then it is a maximally preservative β-removal. (*regularity 2*)

Clearly, *regularity* holds if and only if both *regularity 1* and *regularity 2* hold.

It does *not* follow from *regularity* that if, in our example, $\mathbf{K} \div (\alpha_1 \& \alpha_2 \& \alpha_3)$ is an $\alpha_1 \,\&\, \alpha_2$-removal, then $\mathbf{K} \div (\alpha_1 \& \alpha_2 \& \alpha_3) = \mathbf{K} \div (\alpha_1 \& \alpha_2)$. To see why this would not be plausible as a general principle, let us modify the example and assume that α_2 and α_3 are equally entrenched, whereas α_1 is more entrenched than both α_2 and α_3. Then we should expect $\alpha_3 \notin \mathbf{K} \div (\alpha_1 \& \alpha_2 \& \alpha_3)$ but $\alpha_3 \in \mathbf{K} \div (\alpha_1 \& \alpha_2)$. However, as a limiting case, we can imagine an epistemic agent who is never indifferent between different ways to remove a belief from her belief set. Such an agent will always, when contracting by α, remove one of the smallest removable parts of α. Thus, the contraction by α will itself be the only maximally preservative α-removal. In our example, when contracting by $\alpha_1 \,\&\, \alpha_2 \,\&\, \alpha_3$, this agent will remove exactly one of α_1, α_2, and α_3, and when contracting by $\alpha_1 \& \alpha_2$, she will remove exactly one of α_1 and α_2. Assuming that the relative entrenchment of her beliefs is context-independent, if she removes α_1 when contracting by $\alpha_1 \,\&\, \alpha_2 \,\&\, \alpha_3$, then α_1 is

[6] *Regularity 1* is closely related to Amartya Sen's β property for rational choice behaviour, and *regularity 2* to his α property. (Sen 1970).

also the removed unit of belief when $\alpha_1 \,\&\, \alpha_2$ is contracted. In general, her epistemic behaviour should satisfy the following postulate:

If $\vdash \alpha \rightarrow \beta$ and $\mathbf{K} \div \alpha \not\vdash \beta$, then $\mathbf{K} \div \beta = \mathbf{K} \div \alpha$. (*hyperregularity*)

Hyperregularity implies that for all α and β, $\mathbf{K} \div (\alpha \& \beta) = \mathbf{K} \div \alpha$ or $\mathbf{K} \div (\alpha \& \beta) = \mathbf{K} \div \beta$. This condition has also been called "decomposition" (Alchourrón et al. 1985, p. 525) As was noted by Gärdenfors (1988, p. 66), it is too strong a principle. Thus, *hyperregularity* is a limiting case of some interest, but not a plausible criterion of rational belief change. The same applies to *strong conservativity* and *indecisiveness*, whereas *symmetry*, *conservativity*, and *regularity* are proposed as reasonable postulates for rational belief change.

Axiomatic Characterizations

Symmetry and *conservativity* are sufficient to characterize, together with (G÷1)-(G÷5) and *finitude*, the contractions of belief sets that are generated by partial meet base contraction. Since these are all fairly plausible postulates, this result adds to the credibility of theory contraction through partial meet base contraction.

Theorem 14.1 An operation ÷ on a belief set **K** is generated by partial meet contraction of a finite base for **K** iff ÷ satisfies (G÷1), (G÷2), (G÷3), (G÷4), (G÷5), *finitude*, *symmetry* and *conservativity*.

If *conservativity* is strengthened to the (much less plausible) postulate of *strong conservativity*, then a characterization is obtained of operations that are generated by maxichoice contractions of finite bases.

Theorem 14.2 An operation ÷ on a belief set **K** is generated by maxichoice partial meet contraction of a finite base for **K** iff ÷ satisfies (G÷1), (G÷2), (G÷3), (G÷4), (G÷5), *finitude*, *symmetry*, and *strong conservativity*.

Indecisiveness, in combination with *conservativity*, is a characteristic postulate for operations based on full meet base contraction.

Theorem 14.3 An operation ÷ on a belief set **K** is generated by full meet contraction of a finite base for **K** iff ÷ satisfies (G÷1), (G÷2), (G÷3), (G÷4), (G÷5), *finitude*, *conservativity*, and *indecisiveness*.

Regularity ensures that the partial meet base contraction that generates ÷ is transitively, maximizingly relational (cf. the section "Partial meet contraction").

Theorem 14.4 An operation ÷ on a belief set **K** is generated by transitively, maximizingly relational partial meet contraction of a finite base for **K**, by a completed selection function,[7] iff ÷ satisfies (G÷1), (G÷2), (G÷3), (G÷4), (G÷5), *finitude*, *symmetry*, *conservativity*, and *regularity*.

[7] The completeness of γ is used in the proof. I do not know if it can be dispensed with.

For maxichoice operations that are transitively, maximizingly relational, the following axiomatic characterization has been obtained:

Theorem 14.5 An operation $\div$ on a belief set $\mathbf{K}$ is generated by a transitively, maximizingly relational maxichoice contraction of a finite base for $\mathbf{K}$ iff $\div$ satisfies (G$\div$1), (G$\div$2), (G$\div$3), (G$\div$4), (G$\div$5), *finitude*, *symmetry*, *strong conservativity*, and *hyperregularity*.

Some of the postulates used in Theorems 14.2, 14.3, and 14.5 were shown in the section "The new postulates" to be quite implausible. Indeed, maxichoice and full meet contraction are of interest only as limiting cases. In contrast, Theorems 14.1 and 14.4 only employ fairly plausible postulates of rational belief change. It is proposed that the classes of base-generated contractions of belief sets that are characterized in these theorems represent reasonable types of belief contraction.

Two further important properties have been obtained for the class of operations that were referred to in Theorem 14.4, namely contractions of belief sets that are generated by transitively, maximizingly relational partial meet base contractions:

Theorem 14.6 Let the operation $\div$ on the belief set $\mathbf{K}$ be generated by some transitively, maximizingly relational partial meet contraction of a finite base for $\mathbf{K}$. Then:

(1) If $\mathbf{K} \div \delta \subseteq (\mathbf{K} \div \alpha) \cap (\mathbf{K} \div \beta)$, then $\mathbf{K} \div \delta \subseteq \mathbf{K} \div (\alpha\&\beta)$. (*weak intersection*)
(2) If $\alpha \notin \mathbf{K} \div (\alpha\&\beta)$, then $\mathbf{K} \div (\alpha\&\beta) \subseteq \mathbf{K} \div \alpha$. (*conjunction*)

Weak intersection is a weaker form of Gärdenfors's (G$\div$7) postulate, namely that "the beliefs that are both in $\mathbf{K} \div \alpha$ and $\mathbf{K} \div \beta$ are also in $\mathbf{K} \div (\alpha\&\beta)$" (Gärdenfors 1988, p. 64).[8] It differs from Gärdenfors's original postulate in being restricted to beliefs that are self-sustained in the sense that was accounted for in section "The new postulates". To see that this is a reasonable restriction, let α and β be self-sustained beliefs that have the same degree of entrenchment, and such that $\alpha \vee \beta$ is believed only as a logical consequence of α and β. For a plausible practical example, let α denote that Algiers is a capital and β that Bern is a capital. I have both these beliefs, and they are equally entrenched. If I contract my belief set by α, then β is unperturbed, so that $\beta \in \mathbf{K} \div \alpha$, and as a consequence of that, $\alpha \vee \beta \in \mathbf{K} \div \alpha$. For symmetrical reasons, $\alpha \vee \beta \in \mathbf{K} \div \beta$. However, if I contract my belief set by $\alpha \,\&\, \beta$, then since α and β are equally entrenched I cannot choose between them, so that they must both go. Since neither α nor β is in $\mathbf{K} \div (\alpha\&\beta)$, and $\alpha \vee \beta$ was believed only as a consequence of (each of) these two beliefs, $\alpha \vee \beta$ will be lost as well. Thus, $\alpha \vee \beta \in (\mathbf{K} \div \alpha) \cap (\mathbf{K} \div \beta)$ but $\alpha \vee \beta \notin \mathbf{K} \div (\alpha\&\beta)$, contrary to (G$\div$7) but in accordance with *weak intersection*.

Conjunction is Gärdenfors's (G$\div$8) postulate. To motivate it, the Algiers and Bern example may again be used. In that case, to remove α is a way to remove a specified part of $\alpha\&\beta$. In general, the removal of a part of a certain belief is at least

[8] The formulas of the quotation have been adapted to the notational convention used here.

as economical (conservative, retentive of the original belief set) as the removal of that belief in in entirely. Therefore, if $\mathbf{K} \div (\alpha\&\beta)$ is an α-removal, then $\mathbf{K} \div \alpha$ should be at least as economical as $\mathbf{K} \div (\alpha\&\beta)$.

In conclusion, with base-generated theory contraction we can avoid the problematic *recovery* postulate (G÷6) of the AGM framework, but still have the plausible basic postulates (G÷1)-(G÷5) and the additional postulates (G÷7) (in a weakened but credibilized form) and (G÷8) (in its original form).[9]

Proofs

A set A is *finite-based* iff there is some finite set A' such that $Cn(A') = Cn(A)$. For any non-empty, finite-based set A, $\&A$ denotes the conjunction of all elements of some finite base of A. For any finite, non-empty set A, $\vee(A)$ denotes the disjunction of the elements of A.

The following lemmas will be needed for the proofs:

Lemma 14.1 $\boldsymbol{B} \perp (\alpha\&\beta) \subseteq \boldsymbol{B} \perp \alpha \cup \boldsymbol{B} \perp \beta$.

Proof of Lemma 14.1 Let $W \in \boldsymbol{B} \perp (\alpha\&\beta)$. Then either $W \not\vdash \alpha$ or $W \not\vdash \beta$. It follows from $W \not\vdash \alpha$ and $W \in \boldsymbol{B} \perp (\alpha\&\beta)$ that $W \in \boldsymbol{B} \perp \alpha$, and in the same way from $W \not\vdash \beta$ and $W \in \boldsymbol{B} \perp (\alpha\&\beta)$ that $W \in \boldsymbol{B} \perp \beta$.

Lemma 14.2 If $X \not\subseteq Y \not\subseteq X$ for all $X \in \boldsymbol{B} \perp \alpha$ and $Y \in \boldsymbol{B} \perp \beta$, then $\boldsymbol{B} \perp (\alpha\&\beta) = \boldsymbol{B} \perp \alpha \cup \boldsymbol{B} \perp \beta$.

We will mostly use a special case of the lemma. Namely if $\{X\} = \boldsymbol{B} \perp \alpha$ and $\{Y\} = \boldsymbol{B} \perp \beta$, and $X \not\subseteq Y \not\subseteq X$, then $\{X, Y\} = \boldsymbol{B} \perp (\alpha\&\beta)$.

Proof of Lemma 14.2 One direction follows from *lemma* 14.1. For the other direction, let $X \in \boldsymbol{B} \perp \alpha$. Then $X \not\vdash (\alpha\&\beta)$. In order to prove that $X \in \boldsymbol{B} \perp (\alpha\&\beta)$, suppose to the contrary that there is some W such that $X \subset W \subseteq \boldsymbol{B}$ and $W \not\vdash (\alpha\&\beta)$. From $X \subset W$ it follows that $W \vdash \alpha$. With $W \not\vdash (\alpha\&\beta)$ this yields $W \not\vdash \beta$, from which it follows that $W \subseteq Y$ for some $Y \in \boldsymbol{B} \perp \beta$. We therefore have $X \subset W \subseteq Y$, contradicting the assumption that $X \not\subseteq Y \not\subseteq X$. We may conclude that $X \in \boldsymbol{B} \perp (\alpha\&\beta)$.

In the same way it follows that if $Y \in \boldsymbol{B} \perp \beta$ then $Y \in \boldsymbol{B} \perp (\alpha\&\beta)$.

Lemma 14.3 If $X \in \boldsymbol{B} \perp \alpha$ and $\boldsymbol{B}$ is finite, then there is some β such that $\vdash \alpha \rightarrow \beta$ and $\{X\} = \boldsymbol{B} \perp \beta$.

Proof of Lemma 14.3 If $X = \boldsymbol{B}$, then let $\beta = \alpha$. Otherwise, let $\boldsymbol{B} \backslash X = \{\xi_1, \ldots \xi_n\}$, and let β be $\alpha \vee \xi_1 \vee \cdots \vee \xi_n$. First suppose that $X \vdash \alpha \vee \xi_1 \vee \cdots \vee \xi_n$. It follows from $X \in \boldsymbol{B} \perp \alpha$ that that $X \vdash \xi_k \rightarrow \alpha$ for all $\xi_k \in \boldsymbol{B} \backslash X$. We therefore have $X \vdash$

[9] I do not know if *weak intersection* and *conjunction* can replace *regularity* in theorem 14.4.

α, contrary to the conditions. It may be concluded that $X \nvdash \alpha \vee \xi_1 \vee \cdots \vee \xi_n$. Since $X \cup \{\xi_k\} \vdash \alpha$, and thus $X \cup \{\xi_k\} \vdash \alpha \vee \xi_1 \vee \cdots \vee \xi_n$ holds for all $\xi_k \in \boldsymbol{B}\backslash X$, it follows that $X \in \boldsymbol{B} \perp (\alpha \vee \xi_1 \vee \cdots \vee \xi_n)$.

Next, let $Z \in \boldsymbol{B} \perp (\alpha \vee \xi_1 \vee \cdots \vee \xi_n)$. Since all elements of $\boldsymbol{B}\backslash X$ imply $\alpha \vee \xi_1 \vee \cdots \vee \xi_n$, we have $Z \subseteq X$. From this and $X \in \boldsymbol{B} \perp (\alpha \vee \xi_1 \vee \cdots \vee \xi_n)$ it follows that $\{X\} = \boldsymbol{B} \perp (\alpha \vee \xi_1 \vee \cdots \vee \xi_n)$.

Lemma 14.4 Let $\boldsymbol{B}$ be a finite set. If $\boldsymbol{B} \neq Z \in \boldsymbol{B} \perp \alpha$ for some α, then $\{Z\} = \boldsymbol{B} \perp \mathcal{V}(\boldsymbol{B}\backslash Z)$. ($\mathcal{V}(A)$ is the disjunction of all elements of A.)

Proof of Lemma 14.4 Let $\boldsymbol{B}\backslash Z = \{\xi_1, \ldots \xi_n\}$. It follows by $Z \in \boldsymbol{B} \perp \alpha$ that for each $\xi_k, Z \vdash \xi_k \rightarrow \alpha$. It follows from this and $Z \nvdash \alpha$ that $Z \nvdash (\xi_1 \vee \cdots \vee \xi_n)$ Since every Z' such that $Z \subset Z' \subseteq \boldsymbol{B}$ contains one of $\xi_1, \ldots \xi_n$, we can conclude that $Z \in \boldsymbol{B} \perp (\xi_1 \vee \cdots \vee \xi_n)$.

Next, suppose that $W \in \boldsymbol{B} \perp (\xi_1 \vee \cdots \vee \xi_n)$. Since all elements of $\boldsymbol{B}\backslash Z$ imply $\xi_1 \vee \cdots \vee \xi_n$, $W \cap (\boldsymbol{B}\backslash Z) = \varnothing$, i.e. $W \subseteq Z$. From this and $Z \in \boldsymbol{B} \perp (\xi_1 \vee \cdots \vee \xi_n)$ it follows that $W = Z$. We may conclude that $\{Z\} = \boldsymbol{B} \perp (\xi_1 \vee \cdots \vee \xi_n)$.

Lemma 14.5 Let $\boldsymbol{B}$ be a finite belief base and $\boldsymbol{B}''$ its closure under conjunction. ($\boldsymbol{B}''$ consists of the sentences that are elements of $\boldsymbol{B}$ or conjunctions of elements of $\boldsymbol{B}$.) If an operation $\div$ on $Cn(\boldsymbol{B})$ is generated by a partial meet contraction on $\boldsymbol{B}$, then it is generated by a partial meet contraction on $\boldsymbol{B}''$.

Definitions for the Proof of Lemma 14.5

1. Let A and $\boldsymbol{B}$ be sets of sentences. Then A is $\boldsymbol{B}$-*closed* iff $Cn(A) \cap \mathbf{B} \subseteq A$.
2. $\Re(\boldsymbol{B})$ is the set of $\boldsymbol{B}$-closed subsets of $\boldsymbol{B}$ (Hansson 1991).

Proof of Lemma 14.5 Let f be the function such that for each element A of $\Re(\boldsymbol{B})$, $f(A)$ is the closure under conjunction of A.

We first need to show that f is a one-to-one correspondence. Suppose that it is not. Then there are two elements A and A' of $\Re(\boldsymbol{B})$ such that $A \neq A'$ and $f(A) = f(A')$. We may, without loss of generality, assume that there is then some α such that $\alpha \in A$ and $\alpha \notin A'$. If follows from $\alpha \in A$ and $Cn(A) = Cn(f(A)) = Cn(f(A')) = Cn(A')$ that $\alpha \in Cn(A')$. Since A' is $\boldsymbol{B}$-closed, it follows from $\alpha \in Cn(A')$ and $\alpha \in \boldsymbol{B}$ that $\alpha \in A'$. We may conclude from this contradiction that f is a one-to-one correspondence.

In order to prove the lemma, suppose that the operation $\div$ on $Cn(\boldsymbol{B})$ is based on the partial meet contraction $\sim_\gamma$ on $\boldsymbol{B}$. Let γ'' be the function such that:

(1) If $\boldsymbol{B}'' \perp \alpha \neq \varnothing$, then $\gamma''(\boldsymbol{B}'' \perp \alpha) = \{X \in \boldsymbol{B}'' \perp \alpha \mid f^{-1}(X) \in \gamma(\boldsymbol{B} \perp \alpha)\}$
(2) If $\boldsymbol{B}'' \perp \alpha = \varnothing$, then $\gamma''(\boldsymbol{B}'' \perp \alpha) = \{\boldsymbol{B}''\}$.

We need to prove (1) that γ'' is a selection function for $\boldsymbol{B}''$, and (2) that for all α, $\mathbf{K} \div \alpha = Cn\left(\bigcap \gamma''(\boldsymbol{B}'' \perp \alpha)\right)$.

Part 1: In order to prove that γ'' is a selection function for $\boldsymbol{B}''$ we have to show that if $\boldsymbol{B}'' \perp \alpha \neq \varnothing$, then $\gamma''(\boldsymbol{B}'' \perp \alpha) \neq \varnothing$. Let $\boldsymbol{B}'' \perp \alpha \neq \varnothing$. Then $\boldsymbol{B} \perp \alpha \neq \varnothing$, from which follows that $\gamma(\boldsymbol{B} \perp \alpha)$ is nonempty. Let $X \in \gamma(\boldsymbol{B} \perp \alpha)$. It follows from $Cn(f(X)) = Cn(X)$ that $f(X) \nvdash \alpha$.

Suppose that there is some $Y \subseteq \boldsymbol{B''}$ such that $f(X) \subset Y \nvdash \alpha$. There is then a set Y with this property that is closed under conjunction. It follows that $X \subset f^{-1}(Y) \nvdash \alpha$ and $f^{-1}(Y) \subseteq \boldsymbol{B}$, contrary to $X \in \boldsymbol{B} \perp \alpha$. We may conclude from this contradiction that there is no $Y \subseteq \boldsymbol{B''}$ such that $f(X) \subset Y \nvdash \alpha$. Since $f(X) \nvdash \alpha$ it follows that $f(X) \in \boldsymbol{B''} \perp \alpha$. Since $f^{-1}(f(X)) = X \in \gamma(\boldsymbol{B} \perp \alpha)$, it follows from the construction of γ'' that $\gamma''(\boldsymbol{B''} \perp \alpha) \neq \varnothing$.

Part 2: Since, by the assumptions, $(Cn(\boldsymbol{B})) \div \alpha = Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha))$, it is sufficient to prove that $Cn\left(\bigcap \gamma''(\boldsymbol{B''} \perp \alpha)\right) = Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha))$. Since $\boldsymbol{B''} \perp \alpha \neq \varnothing$ if and only if $\boldsymbol{B} \perp \alpha \neq \varnothing$, and $Cn(\boldsymbol{B''}) = Cn(\boldsymbol{B})$, only the case when $\boldsymbol{B''} \perp \alpha \neq \varnothing$ requires further consideration.

For one direction, let $\delta \in \bigcap \gamma(\boldsymbol{B} \perp \alpha)$. Then $\delta \in Z$ for every $Z \in \gamma(\boldsymbol{B} \perp \alpha)$. By the construction of γ'', $\delta \in Z''$ for every $Z'' \in \gamma''(\boldsymbol{B''} \perp \alpha)$. It follows that $\delta \in \bigcap \gamma''(\boldsymbol{B''} \perp \alpha)$. Thus, $\bigcap \gamma(\boldsymbol{B} \perp \alpha) \subseteq \bigcap \gamma''(\boldsymbol{B''} \perp \alpha)$; from which $Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha)) \subseteq Cn\left(\bigcap \gamma''(\boldsymbol{B''} \perp \alpha)\right)$ can be concluded.

For the other direction, suppose that $\varepsilon \in \bigcap \gamma''(\boldsymbol{B''} \perp \alpha)$. It follows from $\varepsilon \in \boldsymbol{B''}$ that there are elements $\varepsilon_1, \ldots \varepsilon_n$ of $\boldsymbol{B}$ such that $\varepsilon \equiv \varepsilon_1 \& \ldots \& \varepsilon_n$. Let $W \in \gamma(\boldsymbol{B} \perp \alpha)$. By the construction of γ'', $f(W) \in \gamma''(\boldsymbol{B''} \perp \alpha)$. It follows from $f(W) \in \boldsymbol{B''} \perp \alpha$ that $f(\mathrm{W})$ is $\boldsymbol{B''}$-closed. We may conclude from $\{\varepsilon_1, \ldots \& \varepsilon_n\} \subseteq \boldsymbol{B''}$, $\varepsilon \in f(W)$ and the $\boldsymbol{B''}$-closure of $f(\mathrm{W})$ that $\{\varepsilon_1, \ldots \& \varepsilon_n\} \subseteq f(W)$.

It follows from $W \in \boldsymbol{B} \perp \alpha$ that W is $\boldsymbol{B}$-closed. Since $Cn(W) = Cn(f(W))$ and $\{\varepsilon_1, \ldots \& \varepsilon_n\} \subseteq \boldsymbol{B}$, we may conclude from $\{\varepsilon_1, \ldots \& \varepsilon_n\} \subseteq f(W)$ that $\{\varepsilon_1, \ldots \& \varepsilon_n\} \subseteq W$. Since this holds for all $W \in \gamma(\boldsymbol{B} \perp \alpha)$, we have $\{\varepsilon_1, \ldots \& \varepsilon_n\} \subseteq \bigcap \gamma(\boldsymbol{B} \perp \alpha)$. Thus, $\varepsilon \in Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha))$. We have proved that $\bigcap \gamma''(\boldsymbol{B''} \perp \alpha) \subseteq Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha))$, from which $Cn\left(\bigcap \gamma''(\boldsymbol{B''} \perp \alpha)\right) \subseteq Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha))$ follows as desired.

Lemma 14.6 Let the operation $\div$ on the belief set $\mathbf{K}$ be generated by the partial meet contraction $\sim_\gamma$ on the finite base $\boldsymbol{B}$ for $\mathbf{K}$. Then $\mathbf{K} \div \beta$ is a maximally preservative α-removal iff $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \beta) \in \boldsymbol{B} \perp \alpha$.

Proof of Lemma 14.6 For the non-trivial direction, suppose that $\mathbf{K} \div \beta$ is a maximally preservative α-removal and that $\boldsymbol{B} \cap (\mathbf{K} \div \beta)$ is not an element of $\boldsymbol{B} \perp \alpha$. Then there must be some $X \subseteq \boldsymbol{B}$ such that $\boldsymbol{B} \cap (\mathbf{K} \div \beta) \subset X \in \boldsymbol{B} \perp \alpha$ and that there is no δ for which $Cn(X) = \mathbf{K} \div \delta$. However, this is impossible since by *lemma* 14.3 there is some δ such that $\{X\} = \boldsymbol{B} \perp \delta$, and by the definition of partial meet contraction $\boldsymbol{B} \sim_\gamma \delta = X$, so that $\mathbf{K} \div \delta = Cn(X)$.

Lemma 14.7 Let $\div$ be an operator on $\mathbf{K}$ that satisfies *closure* (G$\div$1) and *finitude*, and let $\boldsymbol{B} = \{\&X \mid (\exists \alpha)(X = \mathbf{K} \div \alpha)\}$. Then $Cn(\boldsymbol{B} \cap (\mathbf{K} \div \alpha)) = \mathbf{K} \div \alpha$.

Proof of Lemma 14.7 By the construction, $\&(\mathbf{K} \div \alpha) \in \boldsymbol{B}$. By *closure* (G$\div$1), $\&(\mathbf{K} \div \alpha) \in \mathbf{K} \div \alpha$. It follows that $\&(\mathbf{K} \div \alpha) \in (\boldsymbol{B} \cap (\mathbf{K} \div \alpha))$, so that

$$Cn(\{\&(\mathbf{K} \div \alpha)\}) \subseteq Cn(\boldsymbol{B} \cap (\mathbf{K} \div \alpha)).$$

We also have $\mathbf{K} \div \alpha \subseteq Cn(\{\&(\mathbf{K} \div \alpha)\})$, so that $\mathbf{K} \div \alpha \subseteq Cn(\boldsymbol{B} \cap (\mathbf{K} \div \alpha))$. The other direction follows directly from *closure* (G$\div$1).

Lemma 14.8 Let $\div$ be an operator on the belief set $\mathbf{K}$ that satisfies *closure* (G$\div$1), *success* (G$\div$4), *finitude* and *conservativity*, and let

$$\boldsymbol{B} = \{\&X \mid (\exists\alpha)(X = \mathbf{K} \div \alpha)\}.$$

Then:

$$\text{If } \{X\} = \boldsymbol{B} \perp \delta, \text{ then } X = \boldsymbol{B} \cap (\mathbf{K} \div \delta).$$

Proof of Lemma 14.8 Let $\{X\} = \boldsymbol{B} \perp \delta$. It follows by *success* (G$\div$4) that $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \subseteq X$. Suppose that $X \nsubseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta)$. Then, by the construction of $\boldsymbol{B}$, there is some ϕ such that $\&(\mathbf{K} \div \phi) \in X$ and $\&(\mathbf{K} \div \phi) \notin \boldsymbol{B} \cap (\mathbf{K} \div \delta)$. By *closure* (G$\div$1), $(\mathbf{K} \div \phi) \nsubseteq (\mathbf{K} \div \delta)$.

It follows by *conservativity* that there is some ψ such that $\mathbf{K} \div \delta \subseteq \mathbf{K} \div \psi \nvdash \delta$ and $(\mathbf{K} \div \phi) \cup (\mathbf{K} \div \psi) \vdash \delta$. However, it follows from $\mathbf{K} \div \psi \nvdash \delta$ and $\{X\} = \boldsymbol{B} \perp \delta$ that $\boldsymbol{B} \cap (\mathbf{K} \div \psi) \subseteq X$. Since both $\boldsymbol{B} \cap (\mathbf{K} \div \phi)$ and $\boldsymbol{B} \cap (\mathbf{K} \div \psi)$ are subsets of X, it follows by *lemma* 14.7 that $(\mathbf{K} \div \phi) \cup (\mathbf{K} \div \psi) \nvdash \delta$. We may conclude from this contradiction that $X = \boldsymbol{B} \cap (\mathbf{K} \div \delta)$.

Lemma 14.9 Let $\div$ be an operation on the belief set $\mathbf{K}$ that satisfies *closure* (G$\div$1), *success* (G$\div$4), *finitude*, *symmetry*, *conservativity* and *regularity 2*. Let $\boldsymbol{B} = \{\&X \mid (\exists\alpha)(X = \mathbf{K} \div \alpha)\}$. Then, if $\{X, Y\} \subseteq \boldsymbol{B} \perp \beta$ and $\mathbf{K} \div \beta \subseteq Cn(X)$, there is some β' such that $\{X, Y\} = \boldsymbol{B} \perp \beta'$ and $\mathbf{K} \div \beta' \subseteq Cn(X)$.

Proof of Lemma 14.9 Let $\boldsymbol{B} \perp \beta = \{X, Y, Z_1, \ldots Z_n\}$, and suppose that

$$\mathbf{K} \div \beta \subseteq Cn(X).$$

Then $\boldsymbol{B} \cap (\mathbf{K} \div \beta) \subseteq \boldsymbol{B} \cap Cn(X) = X$.

It follows from $X \in \boldsymbol{B} \perp \beta$, by *lemma* 14.3, that $\{X\} = \boldsymbol{B} \perp \delta$ for some δ such that $\vdash \beta \rightarrow \delta$. By *lemma* 14.8, $X = \boldsymbol{B} \cap (\mathbf{K} \div \delta)$, so that by *closure* (G$\div$1) and *lemma* 14.7, $\mathbf{K} \div \delta = Cn(X)$. Similarly, there are ε and $\zeta_1, \ldots \zeta_n$ such that $\{Y\} = \boldsymbol{B} \perp \varepsilon$ and $\mathbf{K} \div \varepsilon = Cn(Y)$, and that $\{Z_k\} = \boldsymbol{B} \perp \zeta_k$ and $\mathbf{K} \div \zeta_k = Cn(Z_k)$ for all k.

By repeated applications of *lemma* 14.2, $\boldsymbol{B} \perp \beta = \boldsymbol{B} \perp (\delta\&\varepsilon\&\zeta_1\& \ldots \&\zeta_n)$. Thus, for all ϕ, $\boldsymbol{B} \cap (\mathbf{K} \div \phi) \vdash \beta$ iff $\boldsymbol{B} \cap (\mathbf{K} \div \phi) \vdash \delta\&\varepsilon\&\zeta_1\& \ldots \&\zeta_n$. By *lemma* 14.7, $\mathbf{K} \div \phi \vdash \beta$ iff $\mathbf{K} \div \phi \vdash \delta\&\varepsilon\&\zeta_1\& \ldots \&\zeta_n$. By *symmetry*, $\mathbf{K} \div \beta = \mathbf{K} \div (\delta\&\varepsilon\&\zeta_1\& \ldots \&\zeta_n)$.

We have assumed that $\mathbf{K} \div \beta \subseteq Cn(X)$, i.e., $\mathbf{K} \div (\delta\&\varepsilon\&\zeta_1\& \ldots \&\zeta_n) \subseteq \mathbf{K} \div \delta$. Suppose that $\mathbf{K} \div \delta$ is not a maximally preservative $\delta \,\&\, \varepsilon \,\&\, \zeta_1 \,\&\, \ldots \,\&\zeta_n$-removal. Then there is some ϕ such that $\mathbf{K} \div \delta \subset \mathbf{K} \div \phi \nvdash \delta\&\varepsilon\&\zeta_1\& \ldots \&\zeta_n$. By *lemma* 14.7, $Cn(\boldsymbol{B} \cap (\mathbf{K} \div \delta)) \subset Cn(\boldsymbol{B} \cap (\mathbf{K} \div \phi))$, so that

$$\boldsymbol{B} \cap (\mathbf{K} \div \delta)) \subset \boldsymbol{B} \cap (\mathbf{K} \div \phi) \nvdash \delta\&\varepsilon\&\zeta_1\& \ldots \&\zeta_n,$$

contrary to $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \boldsymbol{B} \perp (\delta\&\varepsilon\&\zeta_1\& \ldots \&\zeta_n)$. We may conclude that $\mathbf{K} \div \delta$ is a maximally preservative $\delta \,\&\, \varepsilon \,\&\, \zeta_1 \,\&\, \ldots \,\&\zeta_n$-removal. By *success* (G$\div$4), it is also a $\delta\&\varepsilon$-removal. It follows by *regularity 2* that $\mathbf{K} \div \delta$ is a maximally preservative

δ&ε-removal, so that $\mathbf{K} \div (\delta \div \varepsilon) \subseteq \mathbf{K} \div \delta$. By *lemma* 14.2, $\boldsymbol{B} \perp (\delta \& \varepsilon) = \{X, Y\}$. Since $Cn(X) = \mathbf{K} \div \delta$ we therefore have $\{X, Y\} = \boldsymbol{B} \perp (\delta \& \varepsilon)$ and $\mathbf{K} \div (\delta \& \varepsilon) \subseteq \mathbf{K} \div \delta = Cn(X)$ as desired.

Proof of Theorem 14.1, Left-to-Right Let ÷ be an operation on **K** that is generated by an operator $\sim_\gamma$ of partial meet contraction on a finite belief base $\boldsymbol{B}$ for **K**. It should be obvious that ÷ satifies (G÷1)-(G÷5) and *finitude*.

Symmetry: We use the converse form of *symmetry*. Suppose that $\mathbf{K} \div \alpha \neq \mathbf{K} \div \beta$, i.e., that $Cn\left(\bigcap \gamma(\boldsymbol{B} \perp \alpha)\right) \neq Cn\left(\bigcap \gamma(\boldsymbol{B} \perp \beta)\right)$. Then $\boldsymbol{B} \perp \alpha \neq \boldsymbol{B} \perp \beta$. Without loss of generality we may assume that there is some $X \in \boldsymbol{B} \perp \alpha$ such that $X \notin \boldsymbol{B} \perp \beta$. There are two cases:

Case 1, $X \vdash \beta$: By *lemma* 14.3 there is some δ such that $\{X\} = \boldsymbol{B} \perp \delta$. By the definition of partial meet contraction, $\boldsymbol{B} \sim_\gamma \delta = X$, so that $\mathbf{K} \div \delta = Cn(X)$. It follows that $\mathbf{K} \div \delta \nvdash \alpha$ and $\mathbf{K} \div \delta \vdash \beta$.

Case 2, $X \nvdash \beta$: Then there is some X' such that $X \subset X' \in \boldsymbol{B} \perp \beta$. *Lemma* 14.3 can be used in the same way as in case 1 to show that there is some δ such that $\mathbf{K} \div \delta = Cn(X')$. It follows that $\mathbf{K} \div \delta \nvdash \beta$ and $\mathbf{K} \div \delta \vdash \alpha$.

Conservativity: By *lemma* 14.5, we may assume that $\boldsymbol{B}$ is closed under conjunction.

Suppose that $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$. Then $\& (\mathbf{K} \div \beta) \notin \boldsymbol{B} \sim_\gamma \alpha$ It follows by the definition of partial meet contraction that there is some $X \in \gamma(\boldsymbol{B} \perp \alpha)$ such that $\& (\mathbf{K} \div \beta) \notin X$. By *lemma* 14.3, there is some δ such that $\{X\} = \boldsymbol{B} \perp \delta$. By the definition of partial meet contraction, $X = \boldsymbol{B} \sim_\gamma \delta$. Since $\boldsymbol{B}$ is, by assumption, closed under conjunction, it follows from $\boldsymbol{B} \sim_\gamma \beta \subseteq \boldsymbol{B}$ that $\& \left(\boldsymbol{B} \sim_\gamma \beta\right) \in \boldsymbol{B}$. We also have $\& (\mathbf{K} \div \beta) = \& \left(\boldsymbol{B} \sim_\gamma \beta\right)$, and it therefore follows from $\& (\mathbf{K} \div \beta) \notin \boldsymbol{B} \sim_\gamma \delta \in \boldsymbol{B} \perp \alpha$ that $\left(\boldsymbol{B} \sim_\gamma \delta\right) \cup \{\& (\mathbf{K} \div \beta)\} \vdash \alpha$, from which $(\mathbf{K} \div \delta) \cup (\mathbf{K} \div \beta) \vdash \alpha$ can be concluded. It follows from $\boldsymbol{B} \sim_\gamma \delta \in \gamma(\boldsymbol{B} \perp \alpha)$ that $\boldsymbol{B} \sim_\gamma \alpha \subseteq \boldsymbol{B} \sim_\gamma \delta \nvdash \alpha$, from which we may conclude that $\mathbf{K} \div \alpha \subseteq \mathbf{K} \div \delta \nvdash \alpha$.

Proof of Theorem 14.1, Right-to-Left Let $\boldsymbol{B} = \{\&X | (\exists \alpha)(X = \mathbf{K} \div \alpha)\}$, and let γ be is defined as follows:

(1) If $\boldsymbol{B} \perp \alpha \neq \varnothing$, then $\gamma(\boldsymbol{B} \perp \alpha) = \{X \in \boldsymbol{B} \perp \alpha | (\mathbf{K} \div \alpha) \subseteq Cn(X)\}$.
(2) If $\boldsymbol{B} \perp \alpha = \varnothing$, then $\gamma(\boldsymbol{B} \perp \alpha) = \{\boldsymbol{B}\}$

We need to show (1) that $\boldsymbol{B}$ is a finite base for **K**, (2) that γ is a function, (3) that γ is a selection function for $\boldsymbol{B}$, and (4) that for all α: $\mathbf{K} \div \alpha = Cn\left(\bigcap \gamma(\boldsymbol{B} \perp \alpha)\right)$.

Part 1: It follows from *vacuity* (G÷3) and *finitude* that $\&\mathbf{K} \in \boldsymbol{B}$. Thus, $\mathbf{K} \subseteq Cn(\boldsymbol{B})$. It follows from *inclusion* (G÷2) that $\boldsymbol{B} \subseteq \mathbf{K}$. Thus, $Cn(\boldsymbol{B}) = \mathbf{K}$. By *finitude*, $\boldsymbol{B}$ is finite.

Part 2: Suppose that γ is not a function over the given domain. Then there are α and β such that $\boldsymbol{B} \perp \alpha = \boldsymbol{B} \perp \beta$ and

$$\{X \in \boldsymbol{B} \perp \alpha | (\mathbf{K} \div \alpha) \subseteq Cn(X)\} \neq \{X \in \boldsymbol{B} \perp \alpha | (\mathbf{K} \div \beta) \subseteq Cn(X)\}.$$

It follows that $\mathbf{K} \div \alpha \neq \mathbf{K} \div \beta$. However, from $\boldsymbol{B} \perp \alpha = \boldsymbol{B} \perp \beta$ it follows that for all δ, $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \vdash \alpha$ iff $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \vdash \beta$. By *lemma* 14.7, $Cn(\boldsymbol{B} \cap (\mathbf{K} \div \delta)) = \mathbf{K} \div \delta$.

Thus, $\mathbf{K} \div \delta \vdash \alpha$ iff $\mathbf{K} \div \delta \vdash \beta$. By *symmetry*, $\mathbf{K} \div \alpha = \mathbf{K} \div \beta$, contrary to what was just shown. This contradiction concludes part (2) of the proof.

Part 3: In order to prove that γ is a selection function for $\boldsymbol{B}$, it remains to be shown that if $\boldsymbol{B} \perp \alpha$ is non-empty, then so is $\gamma(\boldsymbol{B} \perp \alpha)$. If $\boldsymbol{B} \perp \alpha$ is non-empty, then α is not a logical truth. By *success* (G÷4), $\mathbf{K} \div \alpha \nvdash \alpha$. Thus $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \nvdash \alpha$, so that there is some X with $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq X \in \boldsymbol{B} \perp \alpha$. By *lemma* 14.7, $Cn(\boldsymbol{B} \cap (\mathbf{K} \div \alpha)) = \mathbf{K} \div \alpha$. It follows that $\mathbf{K} \div \alpha \subseteq Cn(X)$. Then by the definition of γ, $\gamma(\boldsymbol{B} \perp \alpha)$ is nonempty.

Part 4: If α is a logical theorem, then let $\beta \notin \mathbf{K}$. It follows by *vacuity* (G÷3) that $\mathbf{K} \div \beta = \mathbf{K}$. By *conservativity*, if $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$, then there is some δ such that $\mathbf{K} \div \delta \nvdash \alpha$. By *closure* (G÷1), this is impossible. Thus $\mathbf{K} \div \beta \subseteq \mathbf{K} \div \alpha$, i.e. $\mathbf{K} \subseteq \mathbf{K} \div \alpha$. With *inclusion* (G÷2), this yields $\mathbf{K} = \mathbf{K} \div \alpha$. By the definition of partial meet contraction, $\bigcap \gamma(\boldsymbol{B} \perp \alpha) = \boldsymbol{B}$. Using the result of part 1 of the present proof, we obtain $Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha)) = Cn(\boldsymbol{B}) = \mathbf{K} = \mathbf{K} \div \alpha$.

If α is not a logical theorem, then we use the construction of $\boldsymbol{B}$ to obtain $\&(\mathbf{K} \div \alpha) \in \boldsymbol{B}$. By *closure* (G÷1), $\&(\mathbf{K} \div \alpha) \in \mathbf{K} \div \alpha$. The construction of γ yields $\&(\mathbf{K} \div \alpha) \in \bigcap \gamma(\boldsymbol{B} \perp \alpha)$. It follows that $\mathbf{K} \div \alpha \subseteq Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha))$.

For the other direction, suppose that $\varepsilon \notin \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$. If there is no β such that $\varepsilon = \&(\mathbf{K} \div \beta)$, then $\varepsilon \notin \boldsymbol{B}$ so that $\varepsilon \notin \bigcap \gamma(\boldsymbol{B} \perp \alpha)$. If $\varepsilon = \&(\mathbf{K} \div \beta)$ for some β, then it follows from by *closure* (G÷1) from $\&(\mathbf{K} \div \beta) \notin \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$ that $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$. By *conservativity* there is some δ such that $\mathbf{K} \div \alpha \subseteq \mathbf{K} \div \delta \nvdash \alpha$ and $(\mathbf{K} \div \beta) \cup (\mathbf{K} \div \delta) \vdash \alpha$. By *lemma* 14.7, $Cn(\boldsymbol{B} \cap (\mathbf{K} \div \delta)) = \mathbf{K} \div \delta$, so that $(\boldsymbol{B} \cap (\mathbf{K} \div \delta)) \cup \{\&(\mathbf{K} \div \beta)\} \vdash \alpha$. It follows from this and $(\boldsymbol{B} \cap (\mathbf{K} \div \alpha)) \subseteq (\boldsymbol{B} \cap (\mathbf{K} \div \delta)) \nvdash \alpha$ that there is some Y such that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta) \subseteq Y \in \boldsymbol{B} \perp \alpha$ and $\&(\mathbf{K} \div \beta) \notin Y$. By *lemma* 14.7 and the definition of γ, $Y \in \gamma(\boldsymbol{B} \perp \alpha)$. Since $\&(\mathbf{K} \div \beta) \notin Y$, we have $\&(\mathbf{K} \div \beta) \notin \bigcap \gamma(\boldsymbol{B} \perp \alpha)$, i.e. $\varepsilon \notin \bigcap \gamma(\boldsymbol{B} \perp \alpha)$.

Thus if $\varepsilon \notin \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$, then $\varepsilon \notin \bigcap \gamma(\boldsymbol{B} \perp \alpha)$. We may conclude that $\bigcap \gamma(\boldsymbol{B} \perp \alpha) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$. It follows by *lemma* 14.7 that $Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha)) \subseteq \mathbf{K} \div \alpha$.

Proof of Theorem 14.2, Left-to-Right Let $\div$ be an operation on $\mathbf{K}$ that is generated by an operator $\sim_\gamma$ of maxichoice partial meet contraction on a finite belief base $\boldsymbol{B}$ for $\mathbf{K}$. We can make use of the corresponding part of the proof of Theorem 14.1, so that it only remains to prove that *strong conservativity* holds.

By *lemma* 14.5, we may assume that $\boldsymbol{B}$ is closed under conjunction. Suppose that $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$. Then $\boldsymbol{B} \sim_\gamma \beta \not\subseteq \boldsymbol{B} \sim_\gamma \alpha$. It follows that α is not a logical theorem, so that $\boldsymbol{B} \sim_\gamma \alpha \nvdash \alpha$, i.e. $\mathbf{K} \div \alpha \nvdash \alpha$.

Since $\boldsymbol{B}$ is closed under conjunction, $\&(\boldsymbol{B} \sim_\gamma \beta) \in \boldsymbol{B}$. Since $\sim_\gamma$ is maxichoice, $\boldsymbol{B} \sim_\gamma \alpha \in \boldsymbol{B} \perp \alpha$. Since $\&(\boldsymbol{B} \sim_\gamma \beta) \notin \boldsymbol{B} \sim_\gamma \alpha$, we may conclude that $\boldsymbol{B} \sim_\gamma \alpha \cup \{\&(\boldsymbol{B} \sim_\gamma \beta)\} \vdash \alpha$, from which it follows that $\mathbf{K} \div \beta \cup \mathbf{K} \div \alpha \vdash \alpha$.

Proof of Theorem 14.2, Right-to-Left $\boldsymbol{B}$ and γ are constructed in the same way as in the proof of Theorem 14.1. We have to prove: (1) that $\boldsymbol{B}$ is a finite base for $\mathbf{K}$, (2) that γ is a function, (3) that γ is a selection function for $\boldsymbol{B}$, (4) that for all α: $\mathbf{K} \div \alpha = Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha))$, and (5) that γ is maxichoice. Parts 1–3 coincide with the corresponding parts of the proof of Theorem 14.1. Since *strong conservativity*

implies *conservativity*, the proof of part 4 of Theorem 14.1 is also a proof of part 4 of the present proof.

Part 5: Let $\delta \in \boldsymbol{B}\backslash(\boldsymbol{B} \sim_\gamma \alpha)$. By the construction of $\boldsymbol{B}$, $\delta = \&(\mathbf{K} \div \beta)$ for some β. Since $\&(\mathbf{K} \div \beta) \in \boldsymbol{B}$ and $\boldsymbol{B} \sim_\gamma \alpha$ is $\boldsymbol{B}$-closed (cf. the definition for *lemma* 14.5), it follows from $\& (\mathbf{K} \div \beta) \notin \bigcap \gamma (\boldsymbol{B} \perp \alpha)$ that $\& (\mathbf{K} \div \beta) \notin Cn (\bigcap \gamma (\boldsymbol{B} \perp \alpha))$, thus by the result of part 4, $\&(\mathbf{K} \div \beta) \notin \mathbf{K} \div \alpha$. By *closure* (G÷1), $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$. By *strong conservativity*, $\mathbf{K} \div \beta \cup \mathbf{K} \div \alpha \vdash \alpha$. Thus, $\boldsymbol{B} \sim_\gamma \alpha \cup \{\&(\mathbf{K} \div \beta)\} \vdash \alpha$, i.e., $\boldsymbol{B} \sim_\gamma \alpha \cup \{\delta\} \vdash \alpha$. Since this holds for all $\delta \in \boldsymbol{B}\backslash(\boldsymbol{B} \sim_\gamma \alpha)$, we can conclude that $\boldsymbol{B} \sim_\gamma \alpha \in \boldsymbol{B} \perp \alpha$.

Proof of Theorem 14.3, Left-to-Right Let $\div$ be the operation on $\mathbf{K}$ that is generated by the operator $\sim$ of full meet contraction on a finite belief base $\boldsymbol{B}$ for $\mathbf{K}$. Making use of the corresponding part of the proof of Theorem 14.1, it only remains for us to prove that *indecisiveness* holds. Just as in Theorem 14.1 we may, due to *lemma* 14.5, assume that $\boldsymbol{B}$ is closed under conjunction.

Suppose that $\mathbf{K} \div \delta \nvdash \alpha$ and $(\mathbf{K} \div \beta) \cup (\mathbf{K} \div \delta) \vdash \alpha$. Then $\boldsymbol{B} \sim \delta \nvdash \alpha$ and $(\boldsymbol{B} \sim \beta) \cup (\boldsymbol{B} \sim \delta) \vdash \alpha$. Since $\boldsymbol{B} \sim \beta$ and $\boldsymbol{B} \sim \delta$ are both subsets of $\boldsymbol{B}$ there is some subset X of $\boldsymbol{B}$ such that $\boldsymbol{B} \sim \delta \subseteq X \in \boldsymbol{B} \perp \alpha$ and $\boldsymbol{B} \sim \beta \not\subseteq X$. By the definition of full meet contraction, $\boldsymbol{B} \sim \alpha \subseteq X$. Suppose that $\boldsymbol{B} \sim \beta \subseteq \boldsymbol{B} \sim \alpha$. It would then follow from $(\boldsymbol{B} \sim \beta) \cup (\boldsymbol{B} \sim \delta) \vdash \alpha$, $\boldsymbol{B} \sim \delta \subseteq X$ and $\boldsymbol{B} \sim \alpha \subseteq X$ that $X \vdash \alpha$, contrary to $X \in \boldsymbol{B} \perp \alpha$. We may conclude that $\boldsymbol{B} \sim \beta \not\subseteq \boldsymbol{B} \sim \alpha$.

Since all elements of $\boldsymbol{B} \perp \alpha$ are $\boldsymbol{B}$-closed (cf. the definition for *lemma* 14.5), their intersection $\boldsymbol{B} \sim \alpha$ is also $\boldsymbol{B}$-closed. Similarly, $\boldsymbol{B} \sim \beta$ is $\boldsymbol{B}$-closed. It therefore follows from $\boldsymbol{B} \sim \beta \not\subseteq \boldsymbol{B} \sim \alpha$. that $Cn(\boldsymbol{B} \sim \beta) \not\subseteq Cn(\boldsymbol{B} \sim \alpha)$, i.e. $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$.

Proof of Theorem 14.3, Right-to-Left Let $\boldsymbol{B} = \{\&X | (\exists \alpha)(X = \mathbf{K} \div \alpha)\}$. We need to show that (1) $\boldsymbol{B}$ is a finite base for $\mathbf{K}$, (2) for all α, $Cn (\bigcap (\boldsymbol{B} \perp \alpha)) \subseteq \mathbf{K} \div \alpha$, and (3) for all α, $\mathbf{K} \div \alpha \subseteq Cn (\bigcap (\boldsymbol{B} \perp \alpha))$. The proof of part 1 coincides with that of part 1 of Theorem 14.1.

Part 2: We are going to prove that $\bigcap (\boldsymbol{B} \perp \alpha) \subseteq \mathbf{K} \div \alpha$. Let $\zeta \notin \mathbf{K} \div \alpha$. If there is no β such that $\zeta = \mathbf{K} \div \beta$, then it follows by the construction of $\boldsymbol{B}$ that $\zeta \notin \bigcap (\boldsymbol{B} \perp \alpha)$. In the principal case, $\zeta = \& (\mathbf{K} \div \beta)$.

It follows from $\& (\mathbf{K} \div \beta) \notin \mathbf{K} \div \alpha$ and *closure* (G÷1) that $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$. By *conservativity*, there is some δ such that $\mathbf{K} \div \alpha \subseteq \mathbf{K} \div \delta \nvdash \alpha$ and $(\mathbf{K} \div \beta) \cup (\mathbf{K} \div \delta) \vdash \alpha$. By *lemma* 14.7, $(\boldsymbol{B} \cap (\mathbf{K} \div \beta)) \cup (\boldsymbol{B} \cap (\mathbf{K} \div \delta)) \vdash \alpha$. Thus, there is some X such that $\& (\boldsymbol{B} \cap (\mathbf{K} \div \beta)) \notin X \in \boldsymbol{B} \perp \alpha$, i.e. $\& (\mathbf{K} \div \beta) \notin X \in \boldsymbol{B} \perp \alpha$. It follows that $\zeta = \& (\mathbf{K} \div \beta) \notin \bigcap (\boldsymbol{B} \perp \alpha)$.

Thus, if $\zeta \notin \mathbf{K} \div \alpha$, then $\zeta \notin \bigcap (\boldsymbol{B} \perp \alpha)$. i.e. $\bigcap (\boldsymbol{B} \perp \alpha) \subseteq \mathbf{K} \div \alpha$. By *closure* (G÷1), we can conclude that $Cn (\bigcap (\boldsymbol{B} \perp \alpha)) \subseteq \mathbf{K} \div \alpha$.

Part 3: We are going to show that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \bigcap (\boldsymbol{B} \perp \alpha)$. Let $\delta \notin \bigcap (\boldsymbol{B} \perp \alpha)$. If there is no β such that $\delta = \& (\mathbf{K} \div \beta)$, then clearly $\delta \notin \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$. In the principal case, let $\delta = \& (\mathbf{K} \div \beta)$. It follows from $\& (\mathbf{K} \div \beta) \notin \bigcap (\boldsymbol{B} \perp \alpha)$ that there is some X such that $\& (\mathbf{K} \div \beta) \notin X \in \boldsymbol{B} \perp \alpha$.

By *lemma* 14.3, $\{X\} = \boldsymbol{B} \perp \phi$ for some ϕ. By *lemma* 14.8, $X = \boldsymbol{B} \cap (\mathbf{K} \div \phi)$. By *lemma* 14.7, $Cn(X) = \mathbf{K} \div \phi$. Therefore, it follows from $\& (\mathbf{K} \div \beta) \notin \boldsymbol{B} \cap (\mathbf{K} \div \phi) \in \boldsymbol{B} \perp \alpha$ that $\mathbf{K} \div \phi \nvdash \alpha$ and $(\mathbf{K} \div \beta) \cup (\mathbf{K} \div \phi) \vdash \alpha$. By *indecisiveness*,

$\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$. By *closure* (G÷1), $\& (\mathbf{K} \div \beta) \notin \mathbf{K} \div \alpha$, so that $\delta = \& (\mathbf{K} \div \beta) \notin \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$.

We may conclude from this that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \bigcap (\boldsymbol{B} \perp \alpha)$. By *lemma* 14.7, $\mathbf{K} \div \alpha \subseteq Cn (\bigcap (\boldsymbol{B} \perp \alpha))$.

Proof of Theorem 14.4, Left-to-Right Let γ be a complete selection function such that $\sim_\gamma$ is an operator of transitively, maximizingly, relational partial meet contraction on a finite belief base $\boldsymbol{B}$ for $\mathbf{K}$, and that $\sim_\gamma$ generates the operation $\div$ on $\mathbf{K}$. We can make use of the corresponding part of the proof of Theorem 14.1, so that it only remains to prove that *regularity 1* and *regularity 2* hold.

Regularity 1: Suppose that $\vdash \alpha \rightarrow \beta$ and that there is some ζ such that $\mathbf{K} \div \zeta$ is both a β-removal and a maximally preservative α-removal. Let $Z = \boldsymbol{B} \cap (\mathbf{K} \div \zeta)$. By *lemma* 14.6, $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq Z \in \boldsymbol{B} \perp \alpha$.

We are first going to show that $\boldsymbol{B} \cap (\mathbf{K} \div a) = \boldsymbol{B}{\sim_\gamma}\alpha$. It follows from $\mathbf{K} \div \alpha = Cn \left(\boldsymbol{B}{\sim_\gamma}\alpha\right)$ that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) = \boldsymbol{B} \cap Cn \left(\boldsymbol{B}{\sim_\gamma}\alpha\right)$. Since each element of $\gamma (\boldsymbol{B} \perp \alpha)$ is $\boldsymbol{B}$-closed (cf. the definition for *lemma* 14.5), so is $\bigcap \gamma (\boldsymbol{B} \perp \alpha) = \boldsymbol{B}{\sim_\gamma}\alpha$, thus $\boldsymbol{B} \cap Cn \left(\boldsymbol{B}{\sim_\gamma}\alpha\right) = \boldsymbol{B}{\sim_\gamma}\alpha$, i.e. $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) = \boldsymbol{B}{\sim_\gamma}\alpha$.

We now have $\boldsymbol{B}{\sim_\gamma}\alpha \subseteq Z \in \boldsymbol{B} \perp \alpha$ so that, by the completeness of γ, $Z \in \gamma (\boldsymbol{B} \perp \alpha)$. It follows from $Z \nvdash \beta$ and $\vdash \alpha \rightarrow \beta$ that $Z \in \boldsymbol{B} \perp \beta$.

Next, we are going to show that $\gamma (\boldsymbol{B} \perp \beta) \subseteq \gamma (\boldsymbol{B} \perp \alpha)$

Let $X \in \gamma (\boldsymbol{B} \perp \beta)$. Suppose that $X \notin \gamma (\boldsymbol{B} \perp \alpha)$. If $X \in \boldsymbol{B} \perp \alpha$, then $Z \in \gamma (\boldsymbol{B} \perp \alpha)$ yields $X \ll Z$. If $X \notin \boldsymbol{B} \perp \alpha$, then there is some X' such that $X \subset X' \in \boldsymbol{B} \perp \alpha$. It follows by the maximizing property that $X \ll X'$ and by $Z \in \gamma(\boldsymbol{B}\perp\alpha)$ that $X' \not\ll Z$. Transitivity yields $X \ll Z$, in this case as well.

From $X \ll Z$ and $Z \in \boldsymbol{B} \perp \beta$ it follows that $X \notin \gamma (\boldsymbol{B} \perp \beta)$. From this contradiction we may conclude that if $X \in \gamma (\boldsymbol{B} \perp \beta)$ then $X \in \gamma (\boldsymbol{B} \perp \alpha)$, i.e., that $\gamma (\boldsymbol{B} \perp \beta) \subseteq \gamma (\boldsymbol{B} \perp \alpha)$.

Now let $\mathbf{K}\div\delta$ be a maximally preservative β-removal. By *lemma* 14.6,

$$\boldsymbol{B} \cap (\mathbf{K} \div \beta) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \boldsymbol{B} \perp \beta.$$

Since $\boldsymbol{B} \cap (\mathbf{K} \div \beta) = \boldsymbol{B}{\sim_\gamma}\beta$ (that follows in the same way as $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) = \boldsymbol{B}{\sim_\gamma}\alpha$), we have $\boldsymbol{B}{\sim_\gamma}\beta \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \boldsymbol{B} \perp \beta$ so that, by the completeness of γ, $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \gamma (\boldsymbol{B} \perp \beta)$. We have just shown that $\gamma (\boldsymbol{B} \perp \beta) \subseteq \gamma (\boldsymbol{B} \perp \alpha)$, and may conclude that $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \gamma (\boldsymbol{B} \perp \alpha)$.

It also follows from $\gamma (\boldsymbol{B} \perp \beta) \subseteq \gamma (\boldsymbol{B} \perp \alpha)$ that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \beta)$. From this and $\boldsymbol{B}\cap(\mathbf{K} \div \beta) \subseteq \boldsymbol{B}\cap(\mathbf{K} \div \delta)$ it follows that $\boldsymbol{B}\cap(\mathbf{K} \div \alpha) \subseteq \boldsymbol{B}\cap(\mathbf{K} \div \delta)$. We therefore have $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \boldsymbol{B} \perp \alpha$. By *lemma* 14.6, $\mathbf{K}\div\delta$ is a maximally preservative α-removal.

Regularity 2: If β is a logical theorem, then $\mathbf{K}\div\alpha \vdash \beta$ so that *regularity 2* holds vacuously. For the principal case, suppose to the contrary that $\vdash \alpha \rightarrow \beta$, that $\mathbf{K}\div\delta$ is a maximally preservative α-removal, and that it is a β-removal but not a maximally preservative β-removal. It follows by *lemma* 14.6 that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \boldsymbol{B} \perp \alpha$. By $\boldsymbol{B}{\sim_\gamma}\alpha = \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$ (cf. the proof for *regularity 1*) and

the completeness of γ we have $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \gamma\,(\boldsymbol{B} \perp \alpha)$. It follows from $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \in B \perp \alpha$, $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \nvdash \beta$ and $\vdash \alpha \rightarrow \beta$ that $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \boldsymbol{B} \perp \beta$.

Suppose that $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \gamma\,(\boldsymbol{B} \perp \beta)$. Then $\boldsymbol{B} \sim_\gamma \beta \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta)$, i.e. $\boldsymbol{B} \cap (\mathbf{K} \div \beta) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta)$. By *lemma* 14.6, $\boldsymbol{B} \cap (\mathbf{K} \div \beta) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \boldsymbol{B} \perp \beta$ contradicts the condition that $\mathbf{K} \div \delta$ is not a maximally preservative β-removal. We may conclude that $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \notin \gamma\,(\boldsymbol{B} \perp \beta)$. It follows that there is some $Z \in \boldsymbol{B} \perp \beta$ such that $Z \underline{\ll} \boldsymbol{B} \cap (\mathbf{K} \div \delta)$ does not hold.

If $Z \in \boldsymbol{B} \perp \alpha$, then $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \gamma\,(\boldsymbol{B} \perp \alpha)$ yields $Z \underline{\ll} \boldsymbol{B} \cap (\mathbf{K} \div \delta)$. If $Z \notin \boldsymbol{B} \perp \alpha$, then there is some Z' such that $Z \subset Z' \in \boldsymbol{B} \perp \alpha$. The maximizing property yields $Z \ll Z'$. It follows from $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \gamma\,(\boldsymbol{B} \perp \alpha)$ that $Z' \underline{\ll} \boldsymbol{B} \cap (\mathbf{K} \div \delta)$. By transitivity, $Z \ll \boldsymbol{B} \cap (\mathbf{K} \div \delta)$, again contradicting what we have just shown. This contradiction concludes the proof.

Proof of Theorem 14.4, Right-to-Left Let $\boldsymbol{B} = \{\&X | (\exists \alpha)(X = \mathbf{K} \div \alpha)\}$. Let $\underline{\ll}$ be the relation such that $Y \underline{\ll} X$ iff either $Y \subset X$ or there is some β such that $\{X, Y\} \subset \boldsymbol{B} \perp \beta$ and $\mathbf{K} \div \beta \subseteq Cn(X)$. Furthermore, let the function γ be defined as follows:

(1) If $\boldsymbol{B} \perp \alpha \neq \varnothing$, then $\gamma\,(\boldsymbol{B} \perp \alpha) = \{\boldsymbol{B'} \in \boldsymbol{B} \perp \alpha | \boldsymbol{B''} \underline{\ll} \boldsymbol{B'}$ for all $\boldsymbol{B''} \in \boldsymbol{B} \perp \alpha\}$.
(2) Otherwise, $\gamma\,(\boldsymbol{B} \perp \alpha) = \{\boldsymbol{B}\}$.

We need to show that (1) $\boldsymbol{B}$ is a finite base for $\mathbf{K}$, (2) γ is a selection function for $\boldsymbol{B}$, (3) the partial meet contraction $\sim_\gamma$ on $\boldsymbol{B}$ generates the operation $\div$ on $\mathbf{K}$, (4) γ is a completed selection function, and (5) γ is transitively, maximizingly relational by $\underline{\ll}$. The proof of part 1 coincides with that of part 1 of Theorem 14.1.

Part 2: In order to prove that γ is a selection function for $\boldsymbol{B}$, we need to show that if $\boldsymbol{B} \perp \alpha \neq \varnothing$, then $\gamma(\boldsymbol{B} \perp \alpha)$ is non-empty.

It follows by *success* and *inclusion* that there is some $\boldsymbol{B'}$ such that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \boldsymbol{B'} \in \boldsymbol{B} \perp \alpha$. By *lemma* 14.3, there is some δ such that $\{\boldsymbol{B'}\} = \boldsymbol{B} \perp \delta$ and $\vdash \alpha \rightarrow \delta$. By *lemma* 14.8, $\boldsymbol{B'} = \boldsymbol{B} \cap (\mathbf{K} \div \delta)$.

Let $\boldsymbol{B''}$ be any element of $\boldsymbol{B} \perp \alpha$. By *lemma* 14.3, there is some ε such that $\{\boldsymbol{B''}\} = \boldsymbol{B} \perp \varepsilon$ and $\vdash \alpha \rightarrow \varepsilon$. By *lemma* 14.2, $\{\boldsymbol{B'}, \boldsymbol{B''}\} = \boldsymbol{B} \perp (\delta \& \varepsilon)$.

By *lemma* 14.7, it follows from $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \delta) \in \boldsymbol{B} \perp \alpha$ that $\mathbf{K} \div \delta$ is a maximally preservative α-removal. By *success* (G÷4), it is also a $\delta \& \varepsilon$-removal. Since $\vdash \alpha \rightarrow (\delta \& \varepsilon)$ it follows by *regularity 2* that $\mathbf{K} \div \delta$ is a maximally preservative $\delta \& \varepsilon$-removal. Thus $\mathbf{K} \div (\delta \& \varepsilon) \subseteq \mathbf{K} \div \delta$. We therefore have $\{\boldsymbol{B'}, \boldsymbol{B''}\} = \boldsymbol{B} \perp (\delta \& \varepsilon)$ and (by *lemma* 14.7) $\mathbf{K} \div (\delta \& \varepsilon) \subseteq Cn\,(\boldsymbol{B} \cap (\mathbf{K} \div \delta)) = Cn\,(\boldsymbol{B'})$. By the definition of $\underline{\ll}$, $\boldsymbol{B''} \underline{\ll} \boldsymbol{B'}$.

Since this holds for all $\boldsymbol{B''} \in \boldsymbol{B} \perp \alpha$, we may conclude by the construction of γ that $\boldsymbol{B'} \in \gamma\,(\boldsymbol{B} \perp \alpha)$, so that $\gamma(\boldsymbol{B} \perp \alpha)$ is non-empty.

Part 3: By *lemma* 14.7, $Cn\,(\boldsymbol{B} \cap (\mathbf{K} \div \alpha)) = \mathbf{K} \div \alpha$. It is therefore sufficient to prove that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) = \bigcap \gamma\,(\boldsymbol{B} \perp \alpha)$.

If α is a logical theorem, then let $\beta \notin \mathbf{K}$. It follows by *vacuity* (G÷3) that $\mathbf{K} \div \beta = \mathbf{K}$. By *conservativity*, if $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$, then there is some δ such that $\mathbf{K} \div \delta \nvdash \alpha$. By *closure* (G÷1), this is impossible. Thus $\mathbf{K} \div \beta \subseteq \mathbf{K} \div \alpha$, i.e. $\mathbf{K} \subseteq \mathbf{K} \div \alpha$. With *inclusion* (G÷2), this yields $\mathbf{K} = \mathbf{K} \div \alpha$. By the definition of partial meet

contraction, $\bigcap \gamma(\boldsymbol{B} \perp \alpha) = \boldsymbol{B}$. Using the result of part 1 of the present proof, we obtain $Cn(\bigcap \gamma(\boldsymbol{B} \perp \alpha)) = Cn(\boldsymbol{B}) = \mathbf{K} = \mathbf{K} \div \alpha$.

It remains to prove the principal case, in which α is not a logical theorem.

Part 3a: We are going to show that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \bigcap \gamma (\boldsymbol{B} \perp \alpha)$, i.e., that if $X \in \gamma (\boldsymbol{B} \perp \alpha)$, then $\boldsymbol{B} \cap (\mathbf{K} \div a) \subseteq X$.

Let $X \in \gamma (\boldsymbol{B} \perp \alpha)$. By *lemma* 14.3, $\{X\} = B \perp \delta$ for some δ such that $\vdash \alpha \rightarrow \delta$. By *lemma* 14.8, $X = \boldsymbol{B} \cap (\mathbf{K} \div \delta)$ and by *lemma* 14.7, $Cn(X) = \mathbf{K} \div \delta$.

Let $Y_1, \ldots Y_n$ be the elements of $\boldsymbol{B} \perp \alpha$ apart from X. By *lemma* 14.3, for each Y_k there is some ε_k such that $\alpha \rightarrow \varepsilon_k$ and $\{Y_k\} = \boldsymbol{B} \perp \varepsilon_k$. By *lemma* 14.8, $Y_k = \boldsymbol{B} \cap (\mathbf{K} \div \varepsilon_k)$. By *lemma* 14.7, $Cn(Y_k) = \mathbf{K} \div \varepsilon_k$.

By repeated uses of *lemma* 14.2, $\boldsymbol{B} \perp \alpha = \boldsymbol{B} \perp (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n)$. Thus, for all $\phi, \boldsymbol{B} \cap (\mathbf{K} \div \phi) \vdash \alpha$ iff $\boldsymbol{B} \cap (\mathbf{K} \div \phi) \vdash \delta \& \varepsilon_1 \& \ldots \& \varepsilon_n$. By *lemma* 14.7, $\mathbf{K} \div \phi \vdash \alpha$ iff $\mathbf{K} \div \phi \vdash \delta \& \varepsilon_1 \& \ldots \& \varepsilon_n$. By *symmetry*, $\mathbf{K} \div \alpha = \mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n)$.

If $\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \nvdash \delta$, then it follows from $\{X\} = \boldsymbol{B} \perp \delta$ that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) = \boldsymbol{B} \cap (\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n)) \subseteq X$.

In the remaining case, when $\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \vdash \delta$, there is by *success* (G÷4) some ε_k such that $\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \nvdash \varepsilon_k$. Since $\{Yk\} = \boldsymbol{B} \perp \varepsilon_k$,

$$\boldsymbol{B} \cap (\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n)) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \varepsilon_k)\,,$$

thus by *lemma* 14.7, $\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \subseteq \mathbf{K} \div \varepsilon_k$.

We are now going to show that $\mathbf{K} \div \varepsilon_k$ is a maximally preservative $(\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n)$-removal. Suppose that it is not. Then there is some ϕ such that $\mathbf{K} \div \varepsilon_k \subset \mathbf{K} \div \phi \nvdash (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n)$. Clearly, $\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon_k) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \phi)$. Since $\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon_k) = Y_k \in \boldsymbol{B} \perp (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n)$, it follows from

$$\mathbf{K} \div \phi \nvdash (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n)$$

that $\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon_k) = \boldsymbol{B} \cap (\mathbf{K} \div \phi)$. On the other hand, it follows from the construction of $\boldsymbol{B}$ that $\& (\mathbf{K} \div \phi) \in \boldsymbol{B} \cap (\mathbf{K} \div \phi)$, whereas by $\mathbf{K} \div \varepsilon_k \subset \mathbf{K} \div \phi$ we have $\& (\mathbf{K} \div \phi) \notin \boldsymbol{B} \cap (\mathbf{K} \div \varepsilon_k)$, so that $\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon_k) \neq \boldsymbol{B} \cap (\mathbf{K} \div \phi)$. By this contradiction, we can conclude that $\mathbf{K} \div \varepsilon_k$ is a maximally preservative $(\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n)$-removal.

Since $\mathbf{K} \div \varepsilon_k$ is also a $(\delta \& \varepsilon_k)$-removal, it follows by *regularity 1* that all maximally preservative $\delta \& \varepsilon_k$-removals are maximally preservative $\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n$-removals.

We are next going to show that $\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \subseteq \mathbf{K} \div (\delta \& \varepsilon_k)$. Suppose that this is not the case. Then by *conservativity* there is some ψ such that

$$\mathbf{K} \div (\delta \& \varepsilon_k) \subseteq \mathbf{K} \div \psi \nvdash \delta \& \varepsilon_k \text{ and } \mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \cup \mathbf{K} \div \psi \vdash \delta \& \varepsilon_k.$$

Since $\mathbf{K} \div \psi$ is a preservative $\delta \& \varepsilon_k$-removal, there must, by *finitude*, be some maximally preservative $\delta \& \varepsilon_k$-removal $\mathbf{K} \div \psi'$ such that $\mathbf{K} \div \psi \subseteq \mathbf{K} \div \psi'$. Then clearly $\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \cup \mathbf{K} \div \psi' \vdash \delta \& \varepsilon_k$. However, we have just shown that all maximally preservative $\delta \& \varepsilon_k$-removals are maximally preservative $\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n$-removals. Thus $\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \subseteq \mathbf{K} \div \psi'$, contradicting

$$\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \cup \mathbf{K} \div \psi' \vdash \delta \& \varepsilon_k \text{ and } \mathbf{K} \div \psi' \nvdash \delta \& \varepsilon_k.$$

We can conclude that $\mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \subseteq \mathbf{K} \div (\delta \& \varepsilon_k)$.

From the definition of γ, it follows from $Y_k \in \boldsymbol{B} \perp \alpha$ and $X \in \gamma(\boldsymbol{B} \perp \alpha)$ that $Y_k \underline{\ll} X$ and $X \underline{\ll} X$. By the definition of $\underline{\ll}$ and *lemma* 14.9 there is some β such that $\{X, Y_k\} = \boldsymbol{B} \perp \beta$ and $\mathbf{K} \div \beta \subseteq Cn(X) = \mathbf{K} \div \delta$.

By *lemma* 14.2, $\{X, Y_k\} = \boldsymbol{B} \perp (\delta \& \varepsilon_k)$. It follows from $\boldsymbol{B} \perp \beta = \boldsymbol{B} \perp (\delta \& \varepsilon_k)$ that for all ζ, $\boldsymbol{B} \cap (\mathbf{K} \div \zeta) \vdash \beta$ iff $\boldsymbol{B} \cap (\mathbf{K} \div \zeta) \vdash \delta \& \varepsilon_k$. By *lemma* 14.7, $\mathbf{K} \div \zeta \vdash \beta$ iff $\mathbf{K} \div \zeta \vdash \delta \& \varepsilon_k$. By *symmetry*, $\mathbf{K} \div \beta = \mathbf{K} \div (\delta \& \varepsilon_k)$. Thus, $\mathbf{K} \div (\delta \& \varepsilon_k) \subseteq \mathbf{K} \div \delta$. We therefore have $\mathbf{K} \div \alpha = \mathbf{K} \div (\delta \& \varepsilon_1 \& \ldots \& \varepsilon_n) \subseteq \mathbf{K} \div (\delta \& \varepsilon_k) \subseteq \mathbf{K} \div \delta = Cn(X)$, i.e., $\mathbf{K} \div \alpha \subseteq Cn(X)$ so that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \subseteq \boldsymbol{B} \cap (Cn(X)) = X$. This holds for all $X \in \gamma(\boldsymbol{B} \perp \alpha)$, finishing this part of the proof.

Part 3b: In order to show that $\bigcap \gamma(\boldsymbol{B} \perp \alpha) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$, we will assume that $\zeta \notin \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$ and prove that there is some X such that $\zeta \notin X \in \gamma(\boldsymbol{B} \perp \alpha)$.

By the constructions of $\boldsymbol{B}$ and γ, this is trivially true unless $\zeta = \&(\mathbf{K} \div \beta)$ for some β. In that case, it follows by *closure* (G÷1) from $\&(\mathbf{K} \div \beta) \notin \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$ that $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \alpha$. By *conservativity*, there is some δ such that $\mathbf{K} \div \alpha \subseteq \mathbf{K} \div \delta \nvdash \alpha$ and $(\mathbf{K} \div \beta) \cup (\mathbf{K} \div \delta) \vdash \alpha$. It follows by *finitude* that there is some maximally preservative α-removal $\mathbf{K} \div \delta'$ such that $(\mathbf{K} \div \beta) \cup (\mathbf{K} \div \delta') \vdash \alpha$. It follows from $\mathbf{K} \div \delta' \nvdash \alpha$ and $(\mathbf{K} \div \beta) \cup (\mathbf{K} \div \delta') \vdash \alpha$ that $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \delta'$.

Suppose that $\boldsymbol{B} \cap (\mathbf{K} \div \delta') \notin \boldsymbol{B} \perp \alpha$. There is then some W such that $\boldsymbol{B} \cap (\mathbf{K} \div \delta') \subset W \in \boldsymbol{B} \perp \alpha$. By *lemma* 14.3 there is some λ such that $\{W\} = \boldsymbol{B} \perp \lambda$. By *lemma* 14.8, $W = \boldsymbol{B} \cap (\mathbf{K} \div \lambda)$. By *closure* and *lemma* 14.7, $\mathbf{K} \div \delta' \subset \mathbf{K} \div \lambda \vdash \alpha$, contrary to the condition that $\mathbf{K} \div \delta'$ is a maximally preservative α-removal. We may conclude that $\boldsymbol{B} \cap (\mathbf{K} \div \delta') \in \boldsymbol{B} \perp \alpha$.

By *lemma* 14.3, $\{\boldsymbol{B} \cap (\mathbf{K} \div \delta')\} = \boldsymbol{B} \perp \varepsilon$ for some ε such that $\vdash \alpha \rightarrow \varepsilon$, and by *lemma* 14.8 $\boldsymbol{B} \cap (\mathbf{K} \div \delta') = \boldsymbol{B} \cap (\mathbf{K} \div \varepsilon)$ so that by *lemma* 14.7 $\mathbf{K} \div \delta' = \mathbf{K} \div \varepsilon$.

Next, let $Y \in \boldsymbol{B} \perp \alpha$. By *lemmas* 14.3 and 14.8 there is some ϕ such that $\vdash \alpha \rightarrow \phi$, $Y = \boldsymbol{B} \cap (\mathbf{K} \div \phi)$, and $\{Y\} = \boldsymbol{B} \perp \phi$. By *lemma* 14.7, $Cn(Y) = \mathbf{K} \div \phi$. By *lemma* 14.2, $\{\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon), \boldsymbol{B} \cap (\mathbf{K} \div \phi)\} = \boldsymbol{B} \perp (\varepsilon \& \phi)$. We have $\vdash \alpha \rightarrow (\varepsilon \& \phi)$, and $\mathbf{K} \div \varepsilon$ is both a maximally preservative α-removal and, by *success* (G÷4), an $\varepsilon \& \phi$-removal. It follows by *regularity 2* that $\mathbf{K} \div \varepsilon$ is a maximally preservative $\varepsilon \& \phi$-removal, so that $\mathbf{K} \div (\varepsilon \& \phi) \subseteq \mathbf{K} \div \varepsilon$. By the definition of $\underline{\ll}$, $(\boldsymbol{B} \cap (\mathbf{K} \div \phi)) \underline{\ll} (\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon))$. Since this holds for all elements $Y = \boldsymbol{B} \cap (\mathbf{K} \div \phi) \in \boldsymbol{B} \perp \alpha$, it follows by the definition of γ that $(\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon)) \in \gamma(\boldsymbol{B} \perp \alpha)$. We have already shown that $\mathbf{K} \div \beta \not\subseteq \mathbf{K} \div \delta' = \mathbf{K} \div \varepsilon$, so that by *closure* (G÷1), $\&(\mathbf{K} \div \beta) \notin (\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon))$. This concludes this part of the proof.

Part 4: To prove that γ has the completion property, let $\boldsymbol{B} \sim_\gamma \alpha \subseteq X \in \boldsymbol{B} \perp \alpha$. Let $Y \in \boldsymbol{B} \perp \alpha$. Then $\{X, Y\} \subseteq \boldsymbol{B} \perp \alpha$, and by part 3 of the present proof, $\mathbf{K} \div \alpha = Cn(\boldsymbol{B} \sim_\gamma \alpha)$ so that $\mathbf{K} \div \alpha \subseteq Cn(X)$. It follows by the definition of $\underline{\ll}$ that $Y \underline{\ll} X$. Since this holds for all $Y \in \boldsymbol{B} \perp \alpha$, it follows by the definition of γ that $X \in \gamma(\boldsymbol{B} \perp \alpha)$.

Part 5: It follows directly by the construction that γ is relational by $\underline{\ll}$ and that $\underline{\ll}$ has the maximizing property. It remains to be shown that $\underline{\ll}$ is transitive.

We will use the symbol $\leq$ as follows:

$Y \leq X$ iff there is some β such that $\{X, Y\} = \boldsymbol{B} \perp \beta$ and $\mathbf{K} \div \beta \subseteq Cn(X)$.

By *lemma* 14.9, $X \underline{\ll} Y$ iff either $X \subset Y$ or $X \leq Y$. The proof of transitivity can therefore be divided into the following four cases:

(1) If $X \subset Y$ and $Y \subset Z$, then either $X \subset Z$ or $X \leq Z$.
(2) If $X \leq Y$ and $Y \leq Z$, then either $X \subset Z$ or $X \leq Z$.
(3) If $X \subset Y$ and $Y \leq Z$, then either $X \subset Z$ or $X \leq Z$.
(4) If $X \leq Y$ and $Y \subset Z$, then either $X \subset Z$ or $X \leq Z$.

If $X = \boldsymbol{B}$, then $X \not\subset Y$, and $X \leq Y$ implies $X = Y$, so that $Y \underline{\ll} Z$ implies $X \underline{\ll} Z$. If $Y = \boldsymbol{B}$, then $Y \not\subset Z$, and $Y \leq Z$ implies $Y = Z$, so that $X \underline{\ll} Y$ implies $X \underline{\ll} Z$. If $Z = \boldsymbol{B}$, then either $X \subset Z$ or $X = Z$, in both cases yielding $X \underline{\ll} Z$. Thus, in the proofs of the four cases we can assume that $X \neq \boldsymbol{B}$, $Y \neq \boldsymbol{B}$, and $Z \neq \boldsymbol{B}$.

Case 1: Trivial.

Case 2: Suppose that $X \leq Y$ and $Y \leq Z$. By *lemma* 14.4 there are sentences a, b, and c such that $\{X\} = \boldsymbol{B} \perp a$, $\{Y\} = \boldsymbol{B} \perp b$, and $\{Z\} = \boldsymbol{B} \perp c$. By *lemmas* 14.7 and 14.8, $\mathbf{K} \div a = Cn(X)$, $\mathbf{K} \div b = Cn(Y)$, and $\mathbf{K} \div c = Cn(Z)$. By *lemma* 14.2, $\{X, Y\} = \boldsymbol{B} \perp (a \& b)$, and $\{Y, Z\} = \boldsymbol{B} \perp (b \& c)$.

By $Y \leq Z$, there is some β such that $\{Y, Z\} = \boldsymbol{B} \perp \beta$ and $\mathbf{K} \div \beta \subseteq Cn(Z)$. By $\boldsymbol{B} \perp \beta = \boldsymbol{B} \perp (b \& c)$ it follows that for all ε, $\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon) \vdash \beta$ iff $\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon) \vdash (b \& c)$. By *lemma* 14.7, $\mathbf{K} \div \varepsilon \vdash \beta$ iff $\mathbf{K} \div \varepsilon \vdash (b \& c)$. By *symmetry*, $\mathbf{K} \div \beta = \mathbf{K} \div (b \& c)$. Thus we have $\mathbf{K} \div (b \& c) \subseteq Cn(Z)$. By a similar proof it follows that $\mathbf{K} \div (a \& b) \subseteq Cn(Y)$. From $\mathbf{K} \div (a \& b) \subseteq \mathbf{K} \div b$ and $\boldsymbol{B} \cap (\mathbf{K} \div b) = Y \in \boldsymbol{B} \perp (a \& b)$ it follows by *lemma* 14.6 that $\mathbf{K} \div b$ is a maximally preservative $a \& b$-removal.

We are now going to show that $Z \not\subset X$. Suppose to the contrary that $Z \subset X$. Then $\mathbf{K} \div c \subset \mathbf{K} \div a$. It follows from the above construction of a and c by *lemma* 14.4 that $\vdash a \rightarrow c$, so that $\vdash a \& b \rightarrow b \& c$. It follows from $\mathbf{K} \div (a \& b) \subseteq Cn(Y)$ and $Y \in \boldsymbol{B} \perp b$ that $\mathbf{K} \div (a \& b) \nvdash b \& c$. Furthermore, it follows by *lemma* 14.6 from $\mathbf{K} \div (b \& c) \subseteq \mathbf{K} \div c$ and $Z = \boldsymbol{B} \cap (\mathbf{K} \div c) \in \boldsymbol{B} \perp b \& c$ that $\mathbf{K} \div c$ is a maximally preservative $b \& c$-removal. We may then conclude from *regularity 1* (since $\mathbf{K} \div b$ is both a $b \& c$-removal and a maximally preservative $a \& b$-removal) that $\mathbf{K} \div c$ is a maximally preservative $a \& b$-removal, contrary to our assumption that $\mathbf{K} \div c \subset \mathbf{K} \div a \nvdash a \& b$. By this contradiction, we may conclude that $Z \not\subset X$.

Having excluded $Z \subset X$, we have three remaining subcases under case 2: $X \subset Z$, $X = Z$ and $X \not\subseteq Z \not\subseteq X$. If $X \subset Z$, then we are done. If $X = Z$, then we can use $\{X\} = \boldsymbol{B} \perp a$ and $\mathbf{K} \div a \subseteq Cn(X)$ to obtain $X \leq Z$ directly from the definition of $\leq$. The remaining subcase is $X \not\subseteq Z \not\subseteq X$.

When $X \not\subseteq Z \not\subseteq X$ it follows by *lemma* 14.2 that $\{X, Z\} = \boldsymbol{B} \perp (a \& c)$ and $\{X, Y, Z\} = \boldsymbol{B} \perp (a \& b \& c)$. We are first going to prove that $\mathbf{K} \div (a \& b \& c) \subseteq \mathbf{K} \div c$. By *success* (G÷4), $\mathbf{K} \div (a \& b \& c)$ does not imply all three of a, b, and c. If it does not imply a, then it follows from $\{X\} = \boldsymbol{B} \perp a$ that $\&(\mathbf{K} \div (a \& b \& c)) \in X$, so that $\mathbf{K} \div (a \& b \& c) \subseteq Cn(X) = \mathbf{K} \div a$. By similar reasoning for b and c it follows that $\mathbf{K} \div (a \& b \& c)$ is a subset of at least one of $\mathbf{K} \div a$, $\mathbf{K} \div b$, and $\mathbf{K} \div c$.

If $\mathbf{K} \div (a \& b \& c) \subseteq \mathbf{K} \div a$, then by *success* (G÷4), $\mathbf{K} \div (a \& b \& c) \nvdash a \& b$. We have already shown that $\mathbf{K} \div b$ is a maximally preservative $a \& b$-removal, and in the same

way we can show that $\mathbf{K}\div a$ is a maximally preservative $a\&b\&c$-removal. Since it is also an $a\&b$-removal, it follows by *regularity 1* that all maximally preservative $a\&b$-removals are also maximally preservative $a\&b\&c$-removals. Thus, $\mathbf{K}\div b$ is a maximally preservative $a\&b\&c$-removal, so that

$$\mathbf{K} \div (a\&b\&c) \subseteq \mathbf{K} \div b.$$

Thus, if $\mathbf{K} \div (a\&b\&c) \subseteq \mathbf{K} \div a$ then $\mathbf{K} \div (a\&b\&c) \subseteq \mathbf{K} \div b$.

If $\mathbf{K} \div (a\&b\&c) \subseteq \mathbf{K} \div b$, then, by *success* (G÷4), $\mathbf{K}\div(a\&b\&c) \nvdash b\&c$. From $\mathbf{K} \div (b\&c) \subseteq \mathbf{K} \div c$ and $\boldsymbol{B} \cap (\mathbf{K} \div c) = Z \in \boldsymbol{B} \perp (b\&c)$ it follows by *lemma* 14.6 that $\mathbf{K}\div c$ is a maximally preservative $b\&c$-removal. In the same way, it follows that $\mathbf{K}\div b$ is a maximally preservative $a\&b\&c$-removal. Since it is also a $b\&c$-removal, it follows by *regularity 1* that all maximally preservative $b\&c$-removals are maximally preservative $a\&b\&c$-removals. Thus, $\mathbf{K}\div c$ is a maximally preservative $a\&b\&c$-removal, so that $\mathbf{K} \div (a\&b\&c) \subseteq \mathbf{K} \div c$.

Thus, $\mathbf{K}\div(a\&b\&c)$ is a subset of at least one of $\mathbf{K}\div a$, $\mathbf{K}\div b$, and $\mathbf{K}\div c$, and if it is a subset of $\mathbf{K}\div a$ then it is a subset of $\mathbf{K}\div b$ and if it is a subset of $\mathbf{K}\div b$ then it is a subset of $\mathbf{K}\div c$. We may conclude that $\mathbf{K} \div (a\&b\&c) \subseteq \mathbf{K} \div c$.

Since $\boldsymbol{B} \cap (\mathbf{K} \div c) \in \boldsymbol{B} \perp (a\&b\&c)$ it follows by *lemma* 14.6 that $\mathbf{K}\div c$ is a maximally preservative $a\&b\&c$-removal. It is also, by *success* (G÷4), an $a\&c$-removal. It follows by *regularity 2* that $\mathbf{K}\div c$ is a maximally preservative $a\&c$-removal, so that $\mathbf{K} \div (a\&c) \subseteq \mathbf{K} \div c$.

Since $\mathbf{K}\div c = Cn(Z)$ we therefore have $\{X, Z\} = \boldsymbol{B} \perp (a\&c)$ and $\mathbf{K} \div (a\&c) \subseteq Cn(Z)$, so that $X \leq Z$, concluding the proof of case 2.

Case 3: Suppose that $X \subset Y$ and $Y \leq Z$. By *lemma* 14.4 there are a, b, and c such that $\{X\} = \boldsymbol{B} \perp a$, $\{Y\} = \boldsymbol{B} \perp b$, $\{Z\} = \boldsymbol{B} \perp c$, and $\vdash b \rightarrow a$. By *lemmas* 14.7 and 14.8, $\mathbf{K}\div a = Cn(X)$, $\mathbf{K}\div b = Cn(Y)$, and $\mathbf{K}\div c = Cn(Z)$.

We are first going to show that $Z \not\subset X$. If $Z \subset X$, then $Z \subset Y$, contrary to $Y \leq Z$. Thus, $Z \not\subset X$. The subcases $X \subset Z$ and $X = Z$ are treated just as in case 2. It remains to treat the subcase when $X \not\subseteq Z \not\subseteq X$. In that case it follows by *lemma* 14.2 that $\{X, Z\} = \boldsymbol{B} \perp (a\&c)$ and $\{Y, Z\} = \boldsymbol{B} \perp (b\&c)$. In the same way as in case 2, it follows from $Y \leq Z$ that $\mathbf{K} \div (b\&c) \subseteq \mathbf{K} \div c$ and that $\mathbf{K}\div c$ is a maximally preservative $b\&c$-removal.

From $\vdash b \rightarrow a$ it follows that $\vdash b\&c \rightarrow a\&c$. By *success*, $\mathbf{K}\div c$ is an $a\&c$-removal. We can conclude by *regularity 2* that $\mathbf{K}\div c$ is a maximally preservative $a\&c$-removal, so that $\mathbf{K} \div (a\&c) \subseteq \mathbf{K} \div c$. Since $\mathbf{K}\div c = Cn(Z)$ we then have $\{X, Z\} = \boldsymbol{B} \perp (a\&c)$ and $\mathbf{K} \div (a\&c) \subseteq Cn(Z)$, so that $X \leq Z$. This concludes the proof of case 3.

Case 4: Suppose that $X \leq Y \subset Z$. By *lemma* 14.4 there are a, b, and c such that $\{X\} = \boldsymbol{B} \perp a$, $\{Y\} = \boldsymbol{B} \perp b$, $\{Z\} = \boldsymbol{B} \perp c$, and $\vdash c \rightarrow b$. By *lemmas* 14.7 and 14.8 we have $\mathbf{K}\div a = Cn(X)$, $\mathbf{K}\div b = Cn(Y)$, and $\mathbf{K}\div c = Cn(Z)$.

If $Z \subset X$, then $Y \subset X$, contrary to $X \leq Y$. Thus $Z \not\subset X$. The subcases $X = Z$ and $X \subset Z$ are treated as in case 2. In the remaining subcase, $X \not\subseteq Z \not\subseteq X$, it follows by *lemma* 14.2 that $\{X, Y\} = \boldsymbol{B} \perp (a\&b)$ and $\{X, Z\} = \boldsymbol{B} \perp (a\&c)$. By $X \leq Y$, we can show, just as in case 2, that $\mathbf{K} \div (a\&b) \subseteq \mathbf{K} \div b$ and that $\mathbf{K}\div b$ is a maximally preservative $a\&b$-removal.

Suppose that $\mathbf{K} \div (a\&c) \subseteq \mathbf{K} \div a$. Then, by *lemma* 14.6, $\mathbf{K}\div a$ is a maximally preservative $a\&c$-removal. By *success* (G÷4), $\mathbf{K}\div(a\&c) \nvdash a\&b$. From $\vdash c \rightarrow b$ it follows that $\vdash a\&c \rightarrow a\&b$. Since $\mathbf{K}\div a$ is both an $a\&b$-removal and a maximaly preservative $a\&c$-removal, it follows by *regularity 1* that all maximally preservative $a\&b$-removals are maximally preservative $a\&c$-removals. Thus, $\mathbf{K}\div b$ is a maximally preservative $a\&c$-removal. However, it follows from $Y \subset Z \nvdash c$ that $\mathbf{K} \div b \subset \mathbf{K} \div c \nvdash a\&c$, so that $\mathbf{K}\div b$ cannot be a maximally preservative $a\&c$-removal. From this contradiction we may conclude that $\mathbf{K}\div(a\&c) \not\subseteq \mathbf{K}\div a$. By *lemma* 14.7, $\boldsymbol{B} \cap (\mathbf{K} \div (a\&c)) \not\subseteq \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$.

Since $\{\boldsymbol{B} \cap (\mathbf{K} \div a)\} = \boldsymbol{B} \perp a$, it follows from $\boldsymbol{B} \cap (\mathbf{K} \div (a\&c)) \not\subseteq \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$ that $\boldsymbol{B} \cap (\mathbf{K} \div (a\&c)) \vdash a$, so that $\mathbf{K}\div(a\&c) \vdash a$. Therefore, by *success* (G÷4), $\mathbf{K}\div(a\&c) \nvdash c$. Since $\boldsymbol{B} \perp c = \{\boldsymbol{B} \cap (\mathbf{K} \div c)\}$, it follows that $\boldsymbol{B} \cap (\mathbf{K} \div (a\&c)) \subseteq \boldsymbol{B} \cap (\mathbf{K} \div c)$, and by *lemma* 14.7 that $\mathbf{K} \div (a\&c) \subseteq \mathbf{K} \div c$. Since $\mathbf{K}\div c = Cn(Z)$ we then have $\{X, Z\} = \boldsymbol{B} \perp (a\&c)$ and $\mathbf{K} \div (a\&c) \subseteq Cn(Z)$, so that $X \leq Z$. This concludes the proof.

Proof of Theorem 14.5, Left-to-Right Due to the corresponding part of the proof of Theorem 14.2, we only have to show that *hyperregularity* holds.

Suppose that $\vdash \alpha \rightarrow \beta$ and $\mathbf{K}\div\alpha \nvdash \beta$, i.e. $\boldsymbol{B} \sim_\gamma \alpha \nvdash \beta$. By the maxichoice property, $\boldsymbol{B}\sim_\gamma\alpha \in \boldsymbol{B} \perp a$. It follows from $\boldsymbol{B} \sim_\gamma \alpha \nvdash \beta$ that $\boldsymbol{B}\sim_\gamma\alpha \in \boldsymbol{B} \perp \beta$. Let $X \in \boldsymbol{B} \perp \beta$. It follows from $\vdash \alpha \rightarrow \beta$ that there is some X' such that $X \subseteq X' \in \boldsymbol{B} \perp \alpha$. It follows that $X \subseteq X' \underline{\ll} \boldsymbol{B}\sim_\gamma\alpha$. By the maximizing and transitive properties of $\underline{\ll}$, it follows that $X \underline{\ll} \boldsymbol{B} \sim_\gamma \alpha$. Since this holds for all $X \in \boldsymbol{B} \perp \beta$, and $\boldsymbol{B}\sim_\gamma\alpha \in \boldsymbol{B} \perp \beta$, $\boldsymbol{B}\sim_\gamma\alpha \in \gamma(\boldsymbol{B} \perp \beta)$. We may conclude by the maxichoice property that that $\boldsymbol{B} \sim_\gamma \beta = \boldsymbol{B} \sim_\gamma \alpha$, i.e. $\mathbf{K}\div\beta = \mathbf{K}\div\alpha$.

Proof of Theorem 14.5, Right-to-Left Let $\boldsymbol{B} = \{\&X | (\exists\alpha)(X = \mathbf{K} \div \alpha)\}$. Let $\underline{\ll}$ be the relation such that $Y \underline{\ll} X$ iff either $Y \subset X$ or there is some β such that $\{X, Y\} = \boldsymbol{B} \perp \beta$ and $\mathbf{K}\div\beta = Cn(X)$. Let γ be the function such that:

(1) If $\boldsymbol{B} \perp \alpha \neq \varnothing$, then $\gamma(\boldsymbol{B} \perp \alpha) = \{\boldsymbol{B'} \in \boldsymbol{B} \perp \alpha | \boldsymbol{B''} \underline{\ll} \boldsymbol{B'}$ for all $\boldsymbol{B''} \in \boldsymbol{B} \perp \alpha\}$
(2) Otherwise, $\gamma(\boldsymbol{B} \perp \alpha) = \{\boldsymbol{B}\}$

We need to show (1) that $\boldsymbol{B}$ is a finite base for $\mathbf{K}$, (2) that γ is a selection function for $\boldsymbol{B}$, (3) that γ is maxichoice, (4) that the partial meet contraction $\sim_\gamma$ on $\boldsymbol{B}$ generates the operation $\div$ on $\mathbf{K}$, and (5) that γ is transitively, maximizingly relational by $\underline{\ll}$. The proof of part 1 coincides with that of part 1 of Theorem 14.1.

Part 2: To prove that γ is a selection function, it is sufficient to show that if $\boldsymbol{B} \perp \alpha \neq \varnothing$, then $\gamma(\boldsymbol{B} \perp \alpha)$ is non-empty.

We are first going to show that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \in \boldsymbol{B} \perp \alpha$. It follows by *inclusion* (G÷2) and *success* (G÷4) that that $\mathbf{K}\div\alpha$ is a subset of $\mathbf{K}$ that does not imply α. Therefore, $\boldsymbol{B} \cap (\mathbf{K} \div \alpha)$ is a subset of $\boldsymbol{B}$ that does not imply α. Suppose that it is not an element of $\boldsymbol{B} \perp \alpha$. Then there must be some $\phi \in \boldsymbol{B}$ such that $\phi \notin \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$ and $(\boldsymbol{B} \cap (\mathbf{K} \div \alpha)) \cup \{\phi\} \nvdash \alpha$. By the construction of $\boldsymbol{B}$, $\phi = \&(\mathbf{K}\div\delta)$ for some δ. From $\&(\mathbf{K} \div \delta) \notin \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$ it follows by *closure* (G÷1) that $\mathbf{K}\div\delta \not\subseteq \mathbf{K}\div\alpha$, and from $(\boldsymbol{B} \cap (\mathbf{K} \div \alpha)) \cup \{\&(\mathbf{K} \div \delta)\} \nvdash \alpha$ it follows that $(\mathbf{K} \div \alpha) \cup (\mathbf{K} \div \delta)\} \nvdash \alpha$.

This contradicts *strong conservativity*, and we may conclude that (for all α) $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \in \boldsymbol{B} \perp \alpha$.

By *lemma* 14.3 there is some ϕ such that $\vdash \alpha \rightarrow \phi$ and $\{\boldsymbol{B} \cap (\mathbf{K} \div \alpha)\} = \boldsymbol{B} \perp \phi$. By *lemma* 14.7, $\mathbf{K} \div \alpha \nvdash \phi$. Let $Y \in \boldsymbol{B} \perp \alpha$. By *lemma* 14.3 there is some ψ such that $\vdash \alpha \rightarrow \psi$ and $\{Y\} = \boldsymbol{B} \perp \psi$. By *lemma* 14.2, $\{\boldsymbol{B} \cap (\mathbf{K} \div \alpha), Y\} = \boldsymbol{B} \perp (\phi \& \psi)$. Since $\vdash \alpha \rightarrow (\phi \& \psi)$ and $\mathbf{K} \div \alpha \nvdash (\phi \& \psi)$ it follows by *hyperregularity* that $\mathbf{K} \div \alpha = \mathbf{K} \div (\phi \& \psi)$. By *lemma* 14.7 and our definition of $\underline{\ll}$, it follows that $Y \underline{\ll} \boldsymbol{B} \cap (\mathbf{K} \div \alpha)$. Since this holds for all $Y \in \boldsymbol{B} \perp \alpha$, it follows by the definition of γ that $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \in \gamma(\boldsymbol{B} \perp \alpha)$. Thus $\gamma(\boldsymbol{B} \perp \alpha)$ is non-empty whenever $\boldsymbol{B} \perp \alpha$ is non-empty.

Part 3: In order to prove that γ is maxichoice, suppose that it is not. Then there is some α such that there are distinct X and Y with $X, Y \in \gamma(\boldsymbol{B} \perp \alpha)$. It follows, by the definition of γ, that there are ϕ and ψ such that $\{X, Y\} = \boldsymbol{B} \perp \phi = \boldsymbol{B} \perp \psi$, $Cn(X) = \mathbf{K} \div \phi$ and $Cn(Y) = \mathbf{K} \div \psi$.

It follows from $\{X, Y\} = \boldsymbol{B} \perp \phi = \boldsymbol{B} \perp \psi$ that $\{X, Y\} = \boldsymbol{B} \perp (\phi \& \psi)$. By what was shown in part 2 of the present proof, $\boldsymbol{B} \cap (\mathbf{K} \div (\phi \& \psi)) \in \boldsymbol{B} \perp (\phi \& \psi)$. It follows from this and $\boldsymbol{B} \perp \phi = \boldsymbol{B} \perp (\phi \& \psi)$ that $\boldsymbol{B} \cap (\mathbf{K} \div (\phi \& \psi)) \nvdash \phi$. By *lemma* 14.7, $\mathbf{K} \div (\phi \& \psi) \nvdash \phi$. From this it follows by *hyperregularity* that $\mathbf{K} \div (\phi \& \psi) = \mathbf{K} \div \phi$. Similarly, $\mathbf{K} \div (\phi \& \psi) = \mathbf{K} \div \psi$, so that $\mathbf{K} \div \phi = \mathbf{K} \div \psi$. Thus $X = Y$, contrary to our conditions. We may conclude that γ is maxichoice.

Part 4: It was shown in part 2 of the present proof that for all α, $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \in \gamma(\boldsymbol{B} \perp \alpha)$. By part 3, $\{\boldsymbol{B} \cap (\mathbf{K} \div \alpha)\} = \gamma(\boldsymbol{B} \perp \alpha)$. It follows from this, by *lemma* 14.7, that $Cn(\boldsymbol{B} \sim_\gamma \alpha) = \mathbf{K} \div \alpha$.

Part 5: It follows directly by the construction that γ is relational by $\underline{\ll}$ and that the maximizing property holds. In the proof of transitivity, we will use the symbol $\leq$ as follows:

$Y \leq X$ iff there is some β such that $\{X, Y\} = \boldsymbol{B} \perp \beta$ and $\mathbf{K} \div \beta = Cn(X)$.

Thus, $X \underline{\ll} Y$ iff either $X \subset Y$ or $X \leq Y$. The proof of transitivity can therefore be divided into the following four cases:

(1) If $X \subset Y$ and $Y \subset Z$, then either $X \subset Z$ or $X \leq Z$.
(2) If $X \leq Y$ and $Y \leq Z$, then either $X \subset Z$ or $X \leq Z$.
(3) If $X \subset Y$ and $Y \leq Z$, then either $X \subset Z$ or $X \leq Z$.
(4) If $X \leq Y$ and $Y \subset Z$, then either $X \subset Z$ or $X \leq Z$.

If $X = \boldsymbol{B}$, then $X \not\subset Y$, and $X \leq Y$ implies $X = Y$, so that $Y \underline{\ll} Z$ implies $X \underline{\ll} Z$. If $Y = \boldsymbol{B}$, then $Y \not\subset Z$, and $Y \leq Z$ implies $Y = Z$, so that $X \underline{\ll} Y$ implies $X \underline{\ll} Z$. If $Z = \boldsymbol{B}$, then either $X \subset Z$ or $X = Z$, in both cases yielding $X \underline{\ll} Z$. Thus, in the proofs of the four cases we can assume that $X \neq \boldsymbol{B}$, $Y \neq \boldsymbol{B}$, and $Z \neq \boldsymbol{B}$.

Case 1: Trivial.

Case 2: Suppose that $X \leq Y$ and $Y \leq Z$. By *lemma* 14.4, there are sentences a, b, and c such that $\{X\} = \boldsymbol{B} \perp a$, $\{Y\} = \boldsymbol{B} \perp b$, and $\{Z\} = \boldsymbol{B} \perp c$. By *lemma* 14.2 we have $\{X, Y\} = \boldsymbol{B} \perp (a \& b)$ and $\{Y, Z\} = \boldsymbol{B} \perp (b \& c)$

By $X \leq Y$ there is some ϕ such that $\{X, Y\} = \boldsymbol{B} \perp \phi$ and $\mathbf{K} \div \phi = Cn(Y)$. It follows from $\boldsymbol{B} \perp \phi = \boldsymbol{B} \perp (a \& b)$ that for all ε, $\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon) \vdash \phi$ iff $\boldsymbol{B} \cap (\mathbf{K} \div \varepsilon) \vdash (a \& b)$. By *lemma* 14.7, $\mathbf{K} \div \varepsilon \vdash \phi$ iff $\mathbf{K} \div \varepsilon \vdash (a \& b)$. By *symmetry*, $\mathbf{K} \div (a \& b) = \mathbf{K} \div \phi = Cn(Y)$. By a similar proof, $\mathbf{K} \div (b \& c) = Cn(Z)$.

Suppose that $Z \subset X$. By the construction from *lemma* 14.4 that was used above to obtain a, b, and c, $\vdash a \rightarrow c$. It follows that $\vdash a \& b \rightarrow b \& c$. Since $\mathbf{K} \div (a \& b) = Cn(Y)$ and $Y \nvdash b$, we have $\mathbf{K} \div (a \& b) \nvdash b \& c$. By *hyperregularity* we then have $\mathbf{K} \div (a \& b) = \mathbf{K} \div (b \& c)$, i.e., $Cn(Y) = Cn(Z)$, so that $Y = Z$. Then $X \leq Z$ follows directly from $X \leq Y$, and we are done.

If $X \subset Z$, then we are also done. If $X = Z$, then we can use $\{X\} = \boldsymbol{B} \perp \alpha$ and $\boldsymbol{B} \cap (\mathbf{K} \div \alpha) \in \boldsymbol{B} \perp \alpha$, that was obtained in part 2 of the present proof, in the definition of $\leq$, and obtain $X \leq X$.

Finally, we have the case when $Z \not\subseteq X \not\subseteq Z$. Then *lemma* 14.2 yields $\{X, Z\} = \boldsymbol{B} \perp (a \& c)$ and $\{X, Y, Z\} = \boldsymbol{B} \perp (a \& b \& c)$. By *hyperregularity* and $\mathbf{K} \div (a \& b) \nvdash b$ it follows that $\mathbf{K} \div (a \& b) = \mathbf{K} \div b$. Similarly, $\mathbf{K} \div (b \& c) = \mathbf{K} \div c$. By *success* (G÷4), $\mathbf{K} \div (a \& b \& c) \nvdash a$ or $\mathbf{K} \div (a \& b \& c) \nvdash b$ or $\mathbf{K} \div (a \& b \& c) \nvdash c$. By *hyperregularity*, $\mathbf{K} \div (a \& b \& c)$ is identical to one of $\mathbf{K} \div a$, $\mathbf{K} \div b$, and $\mathbf{K} \div c$. Similarly, $\mathbf{K} \div (a \& c)$ is identical to one of $\mathbf{K} \div a$ and $\mathbf{K} \div c$.

Suppose that $\mathbf{K} \div (a \& c) = \mathbf{K} \div a$. It then follows (since, by *hyperregularity*, for all κ and λ, either $\mathbf{K} \div (\kappa \& \lambda) = \mathbf{K} \div \kappa$ or $\mathbf{K} \div (\kappa \& \lambda) = \mathbf{K} \div \lambda$):

(A) Either $\mathbf{K} \div (a \& b \& c) = \mathbf{K} \div (a \& b) = \mathbf{K} \div b$ or $\mathbf{K} \div (a \& b \& c) = \mathbf{K} \div c$.
(B) Either $\mathbf{K} \div (a \& b \& c) = \mathbf{K} \div (a \& c) = \mathbf{K} \div a$ or $\mathbf{K} \div (a \& b \& c) = \mathbf{K} \div b$.
(C) Either $\mathbf{K} \div (a \& b \& c) = \mathbf{K} \div (b \& c) = \mathbf{K} \div c$ or $\mathbf{K} \div (a \& b \& c) = \mathbf{K} \div a$.

Since the three conditions are incompatible, we may conclude that $\mathbf{K} \div (a \& c) \neq \mathbf{K} \div a$. By parts 3 and 4 of the present proof, since $\boldsymbol{B} \perp (a \& c) = \{X, Z\}$, $\mathbf{K} \div (a \& c)$ is either $Cn(X)$ or $Cn(Z)$. In the same way it follows from $\boldsymbol{B} \perp a = \{X\}$ that $\mathbf{K} \div a = Cn(X)$ and from $\boldsymbol{B} \perp c = \{Z\}$ that $\mathbf{K} \div c = Cn(Z)$. Thus, $\mathbf{K} \div (a \& c) = Cn(Z)$. It follows that $X \leq Z$.

Case 3: Suppose that $X \subset Y \leq Z$. By *lemma* 14.4, there are sentences a, b, and c such that $\{X\} = \boldsymbol{B} \perp a$, $\{Y\} = \boldsymbol{B} \perp b$, $\{Z\} = \boldsymbol{B} \perp c$, and $\vdash b \rightarrow a$. $Z \subseteq X$ is impossible, because then $Z \subset Y$, which contradicts $Y \leq Z$. If $X \subset Z$, then we are done. In the remaining case, when $Z \not\subseteq X \not\subseteq Z$, *lemma* 14.2 yields $\{X, Z\} = \boldsymbol{B} \perp (a \& c)$ and $\{Y, Z\} = \boldsymbol{B} \perp (b \& c)$. In the same way as in case 2 we obtain $\mathbf{K} \div (b \& c) = Cn(Z)$. From $\vdash b \rightarrow a$ it follows that $\vdash b \& c \rightarrow a \& c$. Since $Z \nvdash c$ we have $\mathbf{K} \div (b \& c) \nvdash a \& c$. It follows by *hyperregularity* that $\mathbf{K} \div (b \& c) = \mathbf{K} \div (a \& c)$, so that $\mathbf{K} \div (a \& c) = Cn(Z)$. It follows that $X \leq Z$.

Case 4: Suppose that $X \leq Y \subset Z$. By *lemma* 14.4, there are sentences a, b, and c such that $\{X\} = \boldsymbol{B} \perp a, \{Y\} = \boldsymbol{B} \perp b, \{Z\} = \boldsymbol{B} \perp c$, and $\vdash c \rightarrow b$. $Z \subseteq X$ is impossible, because then $Y \subset X$, which contradicts $X \leq Y$. If $X \subset Z$, then we are done. In the remaining case, when $Z \not\subseteq X \not\subseteq Z$, *lemma* 14.2 yields $\{X, Y\} = \boldsymbol{B} \perp (a \& b)$ and $\{X, Z\} = \boldsymbol{B} \perp (a \& c)$. In the same way as in case 2, we obtain $\mathbf{K} \div (a \& b) = Cn(Y)$. We can also prove, in the same way as in case 2, that $\mathbf{K} \div (a \& c)$ is identical to either $\mathbf{K} \div a$ or $\mathbf{K} \div c$.

First, let $\mathbf{K} \div (a \& c) = \mathbf{K} \div a$. From $\vdash c \rightarrow b$ it follows that $\vdash a \& c \rightarrow a \& b$. Since $\mathbf{K} \div a \nvdash a$ we have $\mathbf{K} \div (a \& c) \nvdash a \& b$. It then follows by *hyperregularity* that

$\mathbf{K}\div(a\&c) = \mathbf{K}\div(a\&b)$, so that $\mathbf{K}\div(a\&b) = \mathbf{K}\div a$. By parts 3 and 4 of the present proof, $\mathbf{K}\div a = Cn(X)$. We therefore have $Cn(X) = Cn(Y)$, so that $X = Y$ and $X \subset Z$.

In the other case, when $\mathbf{K}\div(a\&c) = \mathbf{K}\div c$, we use parts 3 and 4 of the present theorem to obtain $\mathbf{K}\div c = Cn(Z)$. It follows from $\{X,Z\} = \boldsymbol{B}\perp(a\&c)$ and $\mathbf{K}\div(a\&c) = Cn(Z)$ that $X \underline{\ll} Z$.

Proof of Theorem 14.6 Let $\boldsymbol{B}$ be a base for $\mathbf{K}$ and γ a selection function for $\boldsymbol{B}$ that is transitively maximizingly relational by $\underline{\ll}$.

PART 1: We are first going to show that $\left(\boldsymbol{B}\sim_\gamma\alpha\right) \cap \left(\boldsymbol{B}\sim_\gamma\beta\right) \subseteq \boldsymbol{B}\sim_\gamma(\alpha\&\beta)$. This is trivial if $\vdash \alpha$ or $\vdash \beta$. For the principal case, in which $\nvdash \alpha$ and $\nvdash \beta$, suppose to the contrary that there is some ζ such that $\zeta \in \boldsymbol{B}\sim_\gamma\alpha$, $\zeta \in \boldsymbol{B}\sim_\gamma\beta$ and $\zeta \notin \boldsymbol{B}\sim_\gamma(\alpha\&\beta)$. It follows from $\zeta \notin \boldsymbol{B}\sim_\gamma(\alpha\&\beta)$ that there is some X such that $\zeta \notin X \in \gamma(\boldsymbol{B} \perp (\alpha\&\beta))$. By *lemma* 14.1, either $X \in \boldsymbol{B} \perp \alpha$ or $X \in \boldsymbol{B} \perp \beta$. Without loss of generality we may assume that $X \in \boldsymbol{B} \perp \alpha$.

Since $\zeta \in \boldsymbol{B}\sim_\gamma\alpha$, $X \notin \gamma(\boldsymbol{B} \perp \alpha)$. Let $Y \in \gamma(\boldsymbol{B} \perp \alpha)$. From $Y \in \boldsymbol{B} \perp \alpha$ it follows that there is some Y' such that $Y \subseteq Y' \in \boldsymbol{B} \perp (\alpha\&\beta)$. If $Y = Y'$ we have $X \ll Y'$ directly, and if $Y \subset Y'$ the maximizing property yields $Y \ll Y'$, which with $X \ll Y$ and transitivity yields $X \ll Y'$. Thus, in both cases $X \ll Y'$, $X \in \gamma(\boldsymbol{B} \perp (\alpha\&\beta))$ and $Y' \in \boldsymbol{B} \perp (\alpha\&\beta)$, contrary to our definitions. This contradiction concludes the proof that $\left(\boldsymbol{B}\sim_\gamma\alpha\right) \cap \left(\boldsymbol{B}\sim_\gamma\beta\right) \subseteq \boldsymbol{B}\sim_\gamma(\alpha\&\beta)$.

Now let $\mathbf{K} \div \delta \subseteq (\mathbf{K} \div \alpha) \cap (\mathbf{K} \div \beta)$. Then $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \subseteq \mathbf{K} \div \alpha$, from which it follows that $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \subseteq \boldsymbol{B}\sim_\gamma\alpha$. Similarly, $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \subseteq \boldsymbol{B}\sim_\gamma\beta$. Since $\left(\boldsymbol{B}\sim_\gamma\alpha\right) \cap \left(\boldsymbol{B}\sim_\gamma\beta\right) \subseteq \boldsymbol{B}\sim_\gamma(\alpha\&\beta)$, we have $\boldsymbol{B} \cap (\mathbf{K} \div \delta) \subseteq \boldsymbol{B}\sim_\gamma(\alpha\&\beta)$, thus $Cn(\boldsymbol{B} \cap (\mathbf{K} \div \delta)) \subseteq Cn\left(\boldsymbol{B}\sim_\gamma(\alpha\&\beta)\right)$, i.e. $\mathbf{K} \div \delta \subseteq \mathbf{K} \div (\alpha\&\beta)$.

PART 2: This is trivial if $\vdash \alpha$, $\vdash \beta$, $\alpha \notin \mathbf{K}$ or $\beta \notin \mathbf{K}$. In the remaining case, suppose that $\alpha \notin \mathbf{K}\div(\alpha\&\beta)$. Then $\alpha \notin \bigcap\gamma(\boldsymbol{B} \perp (\alpha\&\beta))$, and there is some X such that $\alpha \notin X \in \gamma(\boldsymbol{B} \perp (\alpha\&\beta))$. Clearly, $X \in \boldsymbol{B} \perp \alpha$. We are going to prove that $\gamma(\boldsymbol{B} \perp \alpha) \subseteq \gamma(\boldsymbol{B} \perp (\alpha\&\beta))$. Let $Y \in \gamma(\boldsymbol{B} \perp \alpha)$.

We are first going to show that $Y \in \boldsymbol{B} \perp (\alpha\&\beta)$. Suppose to the contrary that $Y \notin \boldsymbol{B} \perp (\alpha\&\beta)$. Then there is some Y' such that $Y \subset Y' \in \boldsymbol{B} \perp (\alpha\&\beta)$. By the maximizing property of $\underline{\ll}$, $Y \ll Y'$. By $X \in \boldsymbol{B} \perp \alpha$ and $Y \in \gamma(\boldsymbol{B} \perp \alpha)$ it follows that $X \underline{\ll} Y$. By transitivity, $X \ll Y'$, which contradicts $X \in \gamma(\boldsymbol{B} \perp (\alpha\&\beta))$ and $Y' \in \boldsymbol{B} \perp (\alpha\&\beta)$. It follows that $Y \in \boldsymbol{B} \perp (\alpha\&\beta)$.

It follows by transitivity from $Y \in \boldsymbol{B} \perp (\alpha\&\beta)$, $X \in \gamma(\boldsymbol{B} \perp (\alpha\&\beta))$ and $X \underline{\ll} Y$ that $Y \in \gamma(\boldsymbol{B} \perp (\alpha\&\beta))$. Thus, $\gamma(\boldsymbol{B} \perp \alpha) \subseteq \gamma(\boldsymbol{B} \perp (\alpha\&\beta))$. This yields $\bigcap(\gamma(\boldsymbol{B} \perp (\alpha\&\beta)) \subseteq \bigcap(\gamma(\boldsymbol{B} \perp \alpha))$ and $\mathbf{K} \div (\alpha\&\beta) \subseteq \mathbf{K} \div \alpha$.

Acknowledgement I would like to thank Peter Gärdenfors, Hans Rott, Wlodek Rabinowicz, and an anonymous referee for valuable comments on an earlier version.

References

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic, 50*, 510–530.

Fuhrmann, A. (1991). Theory contraction through base contraction. *Journal of Philosophical Logic, 20*, 175–203.
Gärdenfors, P. (1984). Epistemic importance and minimal changes of belief. *Australasian Journal of Psychology, 62*, 136–157.
Gärdenfors, P. (1988). *Knowledge in flux. Modeling the dynamics of epistemic states*. Cambridge: MIT Press.
Gärdenfors, P., & Makinson, D. (1988). Revisions of knowledge systems using epistemic entrenchment. In M. Y. Vardi (Ed.), *Proceedings of the second conference on theoretical aspects of reasoning about knowledge* (pp. 83–95).
Hansson, S. O. (1989). New operators for theory change. *Theoria, 55*, 114–132.
Hansson, S. O. (1991). Belief contraction without recovery. *Studia Logica., 50*, 251–260.
Hansson, S. O. (1992). In defense of base contraction. *Synthese, 91*, 239–245.
Hansson, S. O. (1993). Reversing the Levi identity. *Journal of Philosophical Logic, 22*, 637–669.
Makinson, D. (1987). On the status of the postulate of recovery in the logic of theory change. *Journal of Philosophical Logic, 16*, 383–394.
Nebel, B. (1992). Syntax-based approaches to belief revision. In P. Gärdenfors (Ed.), *Belief Revision* (pp. 52–88). Cambridge: Cambridge University Press.
Niederée, R. (1991). Multiple contraction. A further case against Gärdenfors' principle of recovery. In A. Fuhrmann & M. Murreau (Eds.), *The logic of theory change* (Lecture notes in Artificial Intelligence 465, pp. 322–334). Berlin: Springer.
Sen, A. (1970). *Collective choice and social welfare*. San Francisco: Holden-Day, Inc.

Chapter 15
How Infallible but Corrigible Full Belief Is Possible

Isaac Levi

Justifying Changes in Full Belief

Inquirers ought to change beliefs for good reason (Levi 1980 ch. 1, 1991, ch. 1, 2004). What those good reasons are depend on the proximate goals of their inquiries. William James urged us to seek Truth and avoid Error in forming beliefs. He ought to have said: Seek Information and avoid Error. The common features of the proximate goals of scientific inquiries ought to be to answer questions of interest without error and in a manner that yields valuable information.

The beliefs inquirers seek to change are full beliefs. Agent X fully believes that *h* if and only if X is certain that *h* is true. That is to say, X rules out the logical possibility that *h* is false as a serious possibility, takes for granted that *h* is true and uses this information as evidence in efforts to increase the information available to X. Such evidence constitutes the basis for making assessments of credal probability used to evaluate risky choices.

Justifiably changing judgments of credal probability is important to inquiry only insofar as it contributes to the promotion of the goals of inquiry. One does not engage in inquiry in order to justify changes in credal probabilities. In inquiry new error free information is sought. Credal, belief or subjective probabilities are neither true nor false.[1] Changing credal probabilities neither succeeds in nor fails to avoid error. And changing credal probabilities fails to rule out logical possibilities as serious possibilities. So they do not add to information.

[1] As F.P. Ramsey (1990), B. de Finetti (1964) and L.J Savage (1954) rightly observed.

I. Levi (✉)
Columbia University, New York, NY, USA
e-mail: levi@columbia.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_15

The state of credal probability judgment to which X is committed is a function of X's state of full belief or "evidence and background knowledge" and a rule for deriving X's credal probability from X's evidence that I call X's *confirmational commitment* (Levi 1974; 1980). When X's credal state is derivable from X's state of full belief or evidence in accordance with a confirmational commitment, justifying a change in credal state is justifying a change either in X's state of full belief or justifying a change in X's confirmational commitment. As just stated, the proximate aim of a change in X's state of full belief is to avoid importing false belief while acquiring valuable information. Confirmational Commitments, like states of credal probability judgment lack truth values. The proximate aim of modifications of X's confirmational commitment do not involve avoidance of false confirmational commitments or acquiring information by ruling out confirmational commitments as serious possibilities. Changing confirmational commitments is relevant to inquiry aimed at acquiring new and valuable information only insofar as such change facilitate changes in full belief that result in new error free and informative states of full belief.

Philosophical reflection concerning the conditions under which states of credal probability judgment should be modified and how they should be modified is of first rate importance. Credal probability judgment is relevant to the assessment of risks both in practical and theoretical deliberation. But justifying changes in credal probability concerns either the justification of changes in states of full belief or changes in confirmational commitments. Changes in full belief and confirmational commitments alike are justified in terms of how well they promote the acquisition of belief states carrying new, valuable error free information.

Some authors think that in inquiry we seek to change degrees of belief (and disbelief) exhibiting a formal structure similar to Shackle's degrees of belief and surprise (disbelief) (1949; 1961). The goal is to change qualitative beliefs in a sense according to which X believes that h if and only if X's degree of belief reaches some suitable nonnegative threshold short of absolute certainty (Cohen 1970, 1977; Spohn 1988).

Degrees of belief in a satisficing sense and belief that h equivalent to a high satisficing degree of belief (Levi 1967a, 1997 ch. 2.4, 2002) that has reached a certain threshold are suitable as modes of appraisal in inquiry when the inquirer is considering whether to add an item of information to his or her stock of full beliefs. If the initial state of full belief supports a degree of belief that h to a sufficiently high degree relative to the initial state of full belief, the inquirer may be warranted in converting the high degree of belief to a full belief to be added to his or her full beliefs. However, to engage in inquiry or to justify changes in degrees of beliefs or mere beliefs loses its purpose except in a context where the changes in degrees of belief are instrumental to justifying a change in state of full belief.

Change in Doxastic Commitment

Studying changes in states of full belief requires some characterization of the structure of a system of potential states of full belief just as studying changes in the state of a mechanical system requires an understanding of a system of mechanical states. The concern here is, of course, inextricably normative concerned as it is with conditions under which changes from one potential state to another is legitimate and, indeed, justified. Neither classical, statistical nor quantum mechanics is normative in this respect. But the need to consider assumptions about the structure of a "space" of potential states remains the same in the case of changes in states of full belief, changes in classical mechanical states and changes in quantum mechanical states.

In this discussion, the space Ω of potential states of full beliefs is partially ordered by a consequence relation satisfying the conditions of a Boolean algebra. (The algebra is closed under meets and joins of whatever cardinality is required.) Inquirer X is in some state of full belief **K** located in this space. While in the state **K**, X is committed to full belief that each consequence of **K** in the algebra is true. **K** is thus a state of doxastic commitment to the truth of the members of the set of its consequences.

It is also a commitment to a standard for distinguishing the potential states in the algebra that are serious possibilities from those that are not. Because I doubt that there is an algebra that represents all potential states of full belief, I rest content with atomic algebras so that we may postulate a set $\boldsymbol{W}$ of atoms and every potential state in the algebra is the join of a subset of $\boldsymbol{W}$. The state of full belief **K** then distinguishes between the set $\boldsymbol{P}_{\mathbf{K}}$ of atoms that are serious possibilities according to **K** (which are all atoms that are consequences of **K**) and the impossibilities that are ruled out by **K**.

Insofar as a potential state **H** in Ω is representable by a set of sentences S in a regimented language L where the logical consequence relation in L preserves the consequence relation for Ω, **H** is also representable by the set $Cn(S)$ of sentences in L closed under the consequence relation for L. $\boldsymbol{W}$ is representable by a set W of maximally consistent sets of sentences in L and $\boldsymbol{P}_{\mathbf{K}}$ by the set P_{K} of maximally consistent extensions of K.

Thus, if we focus in inquiry on justifying changes in states of full belief we restrict attention to situations where both before and after the change the inquirer's belief state satisfies requirements for logically omniscience. This seems to be absurd. And it is absurd if the inquirer in belief state **K** is burdened with the obligation to recognize on demand all the consequences of his or her full beliefs. I have argued elsewhere for the view that the inquirer may be *committed* to fully believing that h without being able to fulfill the commitment on demand. The inquirer so committed ought to be prepared to acknowledge his or her failure when it is brought to his or her attention and should be prepared to improve his or her performance when opportunity and costs permit. But doxastic changes

that implement such improvements are not changes in doxastic commitment. Even though they play an important role in inquiry, they are not the changes that are the target of inquiry.

The belief changes that are central to inquiry are changes in doxastic commitment. I have suggested for a long time that changes in doxastic commitment that are subject to justification are expansions and contractions.

Expansions are changes from weaker commitments to stronger ones. $\mathbf{K_N}$ is a consequence in the algebra of $\mathbf{K_I}$. The information contained in the initial commitment $\mathbf{K_I}$ is included in the information contained in the new commitment $\mathbf{K_N}$. Contractions are changes from strong commitments to weak one where the inquirer gives up information rather than acquiring it.

Expansions may be described in terms of the information added to an initial state **K**. The expansion of **K** by adding potential state of full belief **H** is $\mathbf{K}\wedge\mathbf{H}$ or $\mathbf{K}^{+}{}_{\mathrm{H}}$. To the extent that potential states are represented by deductively closed sets of sentences in L, the expansion is represented by $Cn(\mathrm{K} \cup \mathrm{H})$. $P_{\mathrm{Cn}(\{\mathrm{K}\cup\mathrm{H}\})} \subseteq P_{\mathrm{K}}$. When H is the set of consequences of a single sentence h, $Cn(\mathrm{K} \cup \{h\}) = \mathrm{K}^{+}{}_{\mathrm{h}}$ or expansion by adding h.

In contraction from **K**, some of the possibilities ruled out by **K** are added to the serious possibilities according to **K**. If **K** has as a consequence $\mathbf{H}^{\mathrm{c}}$, contraction that removes this consequence has $\mathbf{K} \vee \mathbf{H}$ as a consequence. **H** is no longer impossible. But the contraction could be stronger. It could have $\mathbf{K} \vee \mathbf{X}$ as consequence where **X** is stronger than **H**. We shall explore this some more later on.

Many authors take another kind of change in doxastic commitment to be more central than expansion or contraction. Revisions are changes in doxastic commitment where the inquirer adds information to $\mathbf{K_I}$ whether or not the new information is consistent with the information in $\mathbf{K_I}$.[2]

Notice, however, that if **H** is a consequence of $\mathbf{K_I}$ and $\mathbf{K_N}$ has $\mathbf{H}^{\mathrm{c}}$ as a consequence the inquirer starts with full belief that **H** is true and from the perspective to which he is then committed deliberately replaces this conviction with $\mathbf{H}^{\mathrm{c}}$. From the inquirer's initial state of full belief $\mathbf{K_I}$, making the change from $\mathrm{K_I}$ to $\mathbf{K_N}$ is replacing true by false belief. Anyone seeking to avoid importing false belief in making a change in full belief should not deliberately engage in such replacement.

This consideration seems decisive as an objection against recognizing deliberate revision of $\mathbf{K_I}$ in the sense of Alchourrón et al. (1985) as a justifiable form of belief change. The objection is decisive provided the proximate goal of inquiry is to seek new, error free and valuable information at the stage where a change in belief is made. But alternative views of the goals of belief change are available.

(a) One can hold that avoidance of false belief is a desideratum in inquiry where truth and falsity are assessed according to an "external" standard severed from

[2]When the information added is implied by K, the revision is degenerate and identical with **K**. When the information added is consisitent with but not implied by **K**, the revision is a nondegenerate expansion. When the information is inconsistent with **K**, the revision is a *replacement* in the sense of Levi 1980.

any inquirer's point of view or as T.Nagel infamously put it, we consider truth according to the point of view from nowhere.[3] In that case, the inquirer should recognize that **H** in $\mathbf{K}_\mathrm{I}$ might be false as fallibilism requires even though $\mathbf{K}_\mathbf{I}$ is the inquirer's standard for serious possibility.

(b) One can follow John Dewey and Peter Gärdenfors in refusing to consider avoidance of false belief as a desideratum in inquiry.

(c) One can embrace the Messianic Realism of Charles Peirce and Karl Popper and regard convergence on the true complete story of the world as an ultimate goal of inquiry. This Messianic Realism may be supplemented by the thesis that inquirers ought to get closer to the truth as a proximate aim of inquiry as Niiniluoto continues to advocate.

(d) One can reject avoidance of error as a proximate aim of the next change in inquiry but insist that avoidance of error in some finite number of changes after the next change is such a goal.

(e) One can embrace Secular Realism as I and W.V. Quine do and maintain that inquirers ought to avoid error as judged according to the evolving doctrine (i.e., according to $\mathbf{K}_\mathrm{I}$) as a proximate aim at the next change in inquiry. (Niiniluoto 1984).

The secular realist response (e) that I favor argues against the justifiability of revising $\mathbf{K}_\mathrm{I}$ by adding information inconsistent with it in a single step. However, sometimes the net effect of such revision may be justified by justifying each step in a sequence of contractions and expansions.

The disagreement between secular realism and the alternatives is neither a metaphysical nor a semantic one. It is a question of values. Peirce and Popper thought that inquiry should be promoting progress toward the truth at the End of Days. Others might think that progress towards true answers is desirable but avoid the excesses of Messianic Realism. In any case, these approaches can spell out their conceptions of aiming at truth and allow truth and falsity to be judged from some agent's point of view. Cognitive values that should be pursued in inquiry are at issue. Truth and falsehood should be judged by the inquirer from the inquirer's initial state $\mathbf{K}_\mathrm{I}$ of full belief when assessing the options available for change in full belief. Retrospective assessment is from the inquirer's view in state $\mathbf{K}_\mathbf{N}$. And others may judge the truth or falsity of the inquirer's beliefs before and after change from their own belief states. Only assessment of the truth or falsity of the inquirer's beliefs from the point of view from nowhere is incoherent.[4]

In any case, according to Secular Realism (the position I endorsed under the epithet "myopic realism" in Levi 1980) the direct justifiability of replacements is decisively rejectable. Of course, changes of belief state by revision can be

[3] Donald Davidson criticized this view. He argued that inquirers cannot coherently aim at truth (1998). Aiming at truth is, indeed, incoherent if it is judged from the point of view from nowhere rather than as Quine put it from the "evolving doctrine".

[4] If someone insists on the coherence of that point of view, it remains obscure as to the relevance of seeking to avoid error as judged from that point of view.

represented as compositions of sequences of contractions and expansions. And such changes are justified, if each contraction and expansion in such a sequence is justified. I dissent, however, from the popular point of view according to which revision $\mathbf{K}^*_{\mathbf{H}}$ by adding **H** is a response to an input importing **H** in the sense that **K** is the initial state of full belief $\mathbf{K}_{\mathrm{I}}$ and $\mathbf{K}^*_{\mathbf{H}}$ is $\mathbf{K}_{\mathrm{N}}$.

As already noted, the central concern of a discussion of justifiable belief change ought to be changes in full belief rather than changes in degrees of belief understood as degrees of credal probability.[5] I have claimed for a long time that when X fully believes that **H**, X is committed to ruling out the *logical* possibility that **H** is false as a serious possibility. It is incoherent for X to fully believe (be certain) that **H** is true and at the same time to acknowledge that **H** might be false or that there is a small, perhaps infinitesimally small, probability that **H** is false.[6]

Many authors who preach the epistemological doctrine of fallibilism call into question the corrigibility of full beliefs i.e., the justifiable modification of full beliefs once endorsed.[7] Epistemological (or doxastic) infallibilism maintains that when X fully believes (is absolutely certain) that **H**, X rules out the logical possibility that **H** is false as a serious possibility. If there is no serious possibility according to X that *h* is false, it seems puzzling to many how X can justifiably come to doubt that **H**.

[5]When I first advanced this view, I used 'knows' rather than 'fully believes'. I continue to think that *according to X* at time t, X fully believes that *h* if and only if X knows that *h*. That is because I define 'knows' as 'truly believes'. This definition is an expression of the epistemic ideals I have borrowed from the pragmatists who do not require that X justify X's current beliefs but only changes in X's beliefs and the view that truth is judged relative to the evolving doctrine so that according to X everything X fully believes is true. Notice that agent Y can agree that X knows that *h* when Y also fully believes that *h* and will disagree other ways. X and Y will agree that X knows that *h* if and only if X truly believes that *h*.

[6]To fully believe, to be certain, or to know that **H** is not equivalent to judging that the probability that **H** is 1. Setting aside the issue of indeterminacy in probability judgment, subjective or credal probability judgment assigns numerical values in the closed unit interval to potential states of full belief (or to propositions) and their complements when both are judged seriously possible according to **K**. If **H** is fully believed and, hence, a consequence of **K**, it carries probability 1 and is seriously possible while its complement is assigned probability 0 and is ruled out as impossible.

[7]In a well-known example of the conflation of infallibility and incorrigibility, R.C. Jeffrey (1965) argued that reasonable agents should not assign credal probability 1 to propositions because one cannot coherently shift down from probability 1 in conformity with modifying credal probability by Conditionalization. This sort of change in credal state is derivable from the inquirer's state of full belief and the credal probability determined for that state of full belief by the inquirer's credibility function (Carnap 1960) or confirmational commitment provided that the confirmational commitment remains unchanged while the state of full belief is expanded by adding the proposition e and the confirmational commitment satisfies the principle of *confirmational conditionalization* (Levi 1980, 4.3). Confirmational Conditionalization is a synchronic constraint on confirmational commitments. Temporal credal conditionalization is a procedure for changing credal probabilities *if* the state of full belief is expanded by adding a new item *e* of information. That temporal principle does not forbid giving up e. Indeed, as long as the confirmational commitment remains fixed, a change from $\mathbf{K}^+_{\mathbf{e}}$ to K can be derived from confirmational conditionalization as long as the confirmational commitment remains fixed. Jeffrey took for granted that contraction of a state of full belief could not be justified. He did not offer a compelling case for this conclusion.

The puzzle is not premised on a contradiction between saying that X was certain that **H** but no longer is. There is no such contradiction. The issue is whether X can justifiably become uncertain (change from full belief that **H** to doubt as to the truth of **H**) if X is concerned to avoid false belief and maximize the value of information.

The justification should show how an inquirer X concerned to avoid importing false belief while seeking to increase the value of the information available could change views. The two desiderata of avoiding error and acquiring valuable information tend to be in conflict. The more probable X judges **H** to be, the lower the risk of error in coming to believe it and less the value of the information carried by *h* tends to be.

Suppose X initially fully believes that **H**, X may not change belief state, may replace full belief that **H** with full belief that $\mathbf{H}^c$ or move to a position of suspense between **H** and $\mathbf{H}^c$ by contracting from **K** by removing **H**.

From X's initial point of view, remaining with the status quo incurs no risk of false belief.

From the same point of view, replacing full belief that **H** with full belief that $\mathbf{H}^c$ imports false belief deliberately. Replacements are indefensible given the goal of avoiding false belief while increasing the value of information.

Finally moving to a position of suspense between **H** and $\mathbf{H}^c$ by contraction incurs no risk of importing false belief. But the inquirer X will be deliberately giving up information in doing this. And given the goals of inquiry, gratuitous surrender of information looks indefensible.

If not only replacement but contraction also is indefensible given the goals of inquiry, the only kind of change that does not deliberately import error or give up valuable information appears to be expansion. Expansion does to be sure incur a risk of importing false belief. But incurring the risk may be justifiable in a manner compatible with the concern to avoid error as long as the value of the information promised compensates for the risk incurred.

To the extent that expansion can be justified along these lines,[8] the inquirer can overcome the obstacles to giving up information in contraction. Contraction incurs a loss of valuable information. But subsequent inquiry stands some chance of enhancing the informational value of the state of full belief.

Changing from an initial state of full belief $\mathbf{K}_I$ to a contraction $\mathbf{K}_N$ implied by $\mathbf{K}_I$ is entertainably justifiable in at least two contexts:

A. $\mathbf{K}_I$ is inconsistent and contraction amounts to retreat from inconsistency.
B. $\mathbf{K}_I$ is consistent but has a consequence $\mathbf{H}^c$ where **H** is a conjecture that would contribute valuable information to X's store of information were it (counter to the verdict of X's current state of full belief $\mathbf{K}_N$) true. Contraction removing $\mathbf{H}^c$ from K_N is contemplated in order to give a hearing to **H**.

[8] In Levi 1983; 1980, and 1991 I provide an account of the justification of expansion along these lines based on ideas developed in Levi. 1967a and 1967b.

Contraction from Inconsistency

The first puzzle to consider regarding context A is that in order to retreat from inconsistency, the inquirer needs to be in a state of inconsistency. $\mathbf{K}_{\mathrm{I}} = \mathbf{K}_{\perp}$.

An inquirer's state of full belief can be inconsistent in one of two ways. X may be committed to a consistent state of full belief but his doxastic performances may be inconsistent. In that case, removing inconsistency involves either efforts at self-therapy or receiving help from others and their technologies. Indeed, not only must the agent extricate him or herself from inconsistency but must identify his or her consistent state of doxastic commitment. No one can humanly succeed in this endeavor completely. All flesh and blood inquirers are inconsistent in their performances. I have already indicated that I shall not be taking up the question of the considerable therapeutic and engineering tasks involved in realizing the local pockets of fulfillment of doxastic commitment that flesh and blood achieve. In this discussion, the kind of retreat from inconsistency involved in this achievement shall not be considered.

The sort of inconsistent state of full belief of concern here is an inconsistent state of doxastic commitment. The puzzle we need to address is how an inquirer who is rationally fulfilling his or her commitments could end up in such an inconsistent state without involvement in a performance failure? To deliberately expand into inconsistency is never justifiable if the common feature of the proximate goals of all efforts to change doxastic commitments is to obtain valuable new information without importing false beliefs. As long as X's state of full belief is consistent, X is committed to judging the inconsistent state $\mathbf{K}_{\perp}$ to be false. To deliberately expand into inconsistency is to deliberately import false belief into one's evolving doctrine.

The inquirer, however, may deliberately incur a risk of importing false belief, and do so rationally, if some appropriate benefit compensates for the risk. In inquiry where the proximate aim is the acquisition of new and valuable error-free information, making well designed observations is undertaken on the assumption that the beliefs formed in response to the observations made have a good chance of being true and informative. And the inquirer may sometimes be convinced that the testimony of witnesses and experts is reliable and substantive. Consulting external sources of information of both varieties incurs some risk of error. But often such consultation is the only available way of acquiring information relevant to a certain investigation. The value of the information acquired may compensate for the risk of error incurred.

The acquisition of new beliefs via observation and the testimony of experts and witnesses is "direct" in the familiar sense that it is not inference from premises. The consultation involves implementing a program for letting inputs (such as sensory excitations or verbal testimony) determine what information to add to a state of full belief. The sensory inputs and verbal testimonies, however, are not premises from which the inquirer infers a new belief. The addition of a new belief is an outcome of a process initiated by the input. Whatever the psychological details of the process might be, the occurrence of the inputs (the sensory stimulations

or the testimony of the experts or witnesses) do not constitute a change in the inquirer's state of full belief in the doxastic commitment sense. If a change in doxastic commitment occurs, it is a response to that input in conformity with the program being implemented.

Pace Russell and generations of empiricists before and after him, the "data" acquired by such *routine expansion* is directly acquired without being immediately given. Even though the new beliefs are acquired *directly* (i.e., without inference), the acquisition presupposes background information relevant to assessing the reliability of the process of making observations or consulting experts.

Commitment to the program for routine expansion should be distinguished from commitment to implementing it. The inquirer may recognize someone as authoritative on some topic without consulting that agent. X may judge a procedure for making observations reliable without using it. Moreover, the inquirer may be committed in one or both respects without having undertaken the commitments deliberately. However, when undertaking commitment to a program for routine expansion is at issue or when implementing a program to which one is already committed is at issue, modeling the situation as a choice between programs for routine expansion is appropriate.

For purposes of analysis, it is useful to distinguish between the inquirer's point of view $\mathbf{K_{PI}}$ prior to deliberately undertaking commitment to a program for routine expansion and the inquirer's point of view $\mathbf{K}_\mathrm{I}$ when deliberating as to whether to implement the program. We may say that at $\mathbf{K_{PI}}$, the inquirer *precommits* to endorsing the deliverances of an implementation of the program should an implementation be undertaken. At $\mathbf{K}_\mathrm{I}$, the inquirer *commits* to the deliverances of an implementation of the program to which the agent is already committed.

From the inquirer's point of view $\mathbf{K_{PI}}$ implementing a program for routine expansion being contemplated might lead to the formation of beliefs incompatible with $\mathbf{K}_{\mathrm{PI}}$. And when the inquirer is considering implementation of a program already endorsed, implementation might lead to the formation of beliefs incompatible with the inquirer X's current full beliefs in $\mathbf{K}_\mathrm{I}$. Routine expansion has the ability to be *conflict injecting*.

Yet, expansion into inconsistency does not deliberately add false belief to $\mathbf{K}_\mathrm{I}$. The inquirer X has risked importing error for the sake of the information to be acquired and has lost the gamble. The outcome is that X has expanded into inconsistency $\mathbf{K}_\perp$ from $\mathbf{K}_\mathrm{I}$.[9]

Inadvertent expansion into inconsistency is a byproduct of the need to use routine expansion to acquire new information that, perhaps, could not be obtained at the time by other means. This kind of inconsistency does not represent failure to fulfill

[9]Regardless of whether the implementation is the product of habitual or customary practice or is deliberate, the inquirer is committed to implementing the program prior to its implementation. The inquirer is precommitted. The inquirer is (pre) committed to the results of implementing the routine regardless of what they might be. When the result is expansion into inconsistency, the importation of false belief is inadvertent.

commitment. To the contrary, it fulfills a commitment to follow a program of routine expansion regardless of where it might lead.

$K_\perp$ fails as a standard for serious possibility and as a resource for specifying truth conditions for hypotheses in a sense of truth according to which an inquirer should be seeking to avoid false belief. It seems obvious that inquirers should retreat from such inconsistency – i.e., contract from $\mathbf{K}_\perp$.

Contraction from inconsistency cannot be justified as the choice of the best option among the (epistemic) options available. Such justification should be based on the inquirer's point of view prior to choice. Prior to choice, however, the inquirer's point of view is inconsistent. An inconsistent state of full belief cannot be used coherently as a standard for serious possibility, for judging truth or as evidence. In Levi 1980, footnote pp. 59–60, I worried about this question of coherence. I suggested treating the language in which the inconsistent state of full belief is expressed syntactically and rationalizing changes in a consistent metalanguage. Olsson (2003) rightly took me to task for not living up to my own commitments concerning the characterization of the state of full belief as a standard for serious possibility. Olsson's objection seems unanswerable according to the approach I favor. But according to that approach, his proposed cure is worse than the disease. Olsson maintained that there are admissible programs for routine expansion that avoid expansion into inconsistency and that programs for routine expansion should be restricted to these. In Levi 2003, I showed that routine expansion is inescapably conflict injecting – counter to what Olsson claimed. The only way to avoid routine expansion into inconsistency is to avoid routine expansion altogether. This remedy seemed to be as unacceptable to Olsson as it is to me. I concluded that we need to consider other ways to address the predicament of how to justify retreat from inconsistency.

In Levi 2003, I suggested that a program for routine expansion should be accompanied by a precommitment to a plan providing a response in case implementation of the program leads to expansion into inconsistency. From the point of view $\mathbf{K}_{\mathrm{PI}}$ the inquirer can coherently identify a set of appropriate contractions from inconsistency and evaluate their merits.

Programs for routine expansion provide for expansions that yield many competing answers to a given question. As a consequence, expansion into inconsistency can take on a diversity of forms depending on the item of information h added to $\mathbf{K_I}$ that yields inconsistency. Given a specific h, the plan stipulates that one of the following contractions from $\mathbf{K}_\perp$ is implemented:

(i) Return to $\mathbf{K_I}$. The expansion prescribed by the routine is nullified.
(ii) Replacement of the consequence $\mathbf{H^c}$ of $\mathbf{K}_\mathrm{I}$ by $\mathbf{H}$. The contraction from inconsistency recommended is representable as an expansion by adding $\mathbf{H}$ to a contraction of $\mathbf{K}_\mathrm{I}$ by removing $\mathbf{H}^\mathrm{c}$.[10]

[10]The equivalence of a replacement to an expansion of a contraction is the *Commensuration* Thesis. (Levi 1991, p.65) The Commensuration thesis is trivial as long the domain of potential states of full belief is partially ordered in a manner satisfying the requirements of a Boolean algebra.

(iii) Contraction of $\mathbf{K}_\mathbf{I}$ by removing $\mathbf{H}^c$.

Each of the three plans recommends returning the state of full belief to a potential state that is formally representable as a transformation of $\mathbf{K}_\mathbf{I}$. The important point to be emphasized is that, nonetheless, each if the three plans implements a potential contraction from inconsistency.

Exercising option (i) is contracting from $\mathbf{K}_\perp$ in a way that returns the state of full belief to the *status quo ex ante*.

Exercising option (ii) is also a move to a consistent potential state of full belief from $\mathbf{K}_\perp$. It is not a move from $\mathbf{K}_\mathbf{I}$ to a replacement of $\sim h$ by h in $\mathbf{K}_\mathbf{I}$. It is a move from $\mathbf{K}_\perp$ to the aforementioned replacement.

Similarly, option (iii) is a move from inconsistency to a contraction of $\mathbf{K}_\mathbf{I}$ by removing $\mathbf{H}^c$.

We should look at (i), (ii), (iii) as three optional precommitment plans for retreating from inconsistency. The three alternatives may be formally representable as transformations of $\mathbf{K}_\mathbf{I}$. But the challenge that provokes consideration of them is an expansion of $\mathbf{K}_\mathbf{I}$ into inconsistency. The task to which all three are responsive is how to retreat from inconsistency.

Thus, even though retreating from $\mathbf{K}_\perp$ is an option as a precommitment plan, remaining with $\mathbf{K}_\mathbf{I}$ is not an option. Similarly replacement of $\mathbf{H}^c$ in $\mathbf{K}_\mathbf{I}$ by $\mathbf{H}$ is not an option even though shifting from inconsistency to the replacement is. Similar remarks apply to contraction of $\mathbf{K}_\mathbf{I}$ by removing *h.*

Formally, replacement of $\mathbf{H}^c$ by $\mathbf{H}$ in $\mathbf{K}_\mathbf{I}$ is equivalent to an expansion by adding $\mathbf{H}$ to a contraction by removing $\mathbf{H}^c$ from $\mathbf{K}_\mathbf{I}$. But as is well known, there are many contractions of $\mathbf{K}_\mathbf{I}$ that remove $\mathbf{H}^c$. We need some way of recommending which contractions removing $\mathbf{H}^c$ from $\mathbf{K}_\mathbf{I}$ can serve as ingredients in contractions from $\mathbf{K}_\mathbf{I}$ in precommitment plans of the three types mentioned above.

Deliberation aimed at adopting a precommitment plan at $\mathbf{K}_\mathbf{PI}$ compares the admissible options of type (i), (ii) and (iii) and determines best or admissible precommitment plans from among these. Option of type (i) is formally the same as first contracting by removing $\mathbf{H}^c$ and then expanding by adding $\mathbf{H}^c$. The type (ii) option (replacement of $\mathbf{H}^c$ in $\mathbf{K}_\mathbf{I}$ by $\mathbf{H}$) is formally equivalent to an expansion by adding $\mathbf{H}$ to a contraction of $\mathbf{K}_\mathbf{I}$ by removing $\mathbf{H}^c$.

Both (i) and (ii), therefore, are representable formally as an expansion of the potential state of full belief that is the outcome of contraction type (iii) where the inquirer is in suspense with respect to the truth of $\mathbf{H}$. The upshot is that the third option yields a belief state that carries less information and, hence, no more informational value than the other two options.

Suppose that result of implementing (i) is a belief state carrying more (less) informational value than the belief state resulting from implementing (ii). On the assumption (that I favor) according to which the inquirer should seek to minimize loss of informational value, if the inquirer should have to retreat from inconsistency due to having expanded $\mathbf{K}_\mathbf{I}$ by adding $\mathbf{H}^c$, the inquirer should follow option (i) or option (ii).

On the other hand, if options (i) and (ii) result in belief states carrying equal informational value or if they are noncomparable with respect to informational value, the inquirer should follow option (iii). This recommendation is based on the assumption that the informational value of the join of two potential states of full belief is the minimum informational value carried by the pair of states. This is the assumption I have used in evaluating loss of damped informational value in contraction in Levi 2004 and in Levi 1991 and 1997.

On this view, neither information acquired via observation nor from expert witnesses is categorically authoritative. If it were, option (ii) would be mandatory in all cases. All but the most rabid empiricists agree that the adoption of (ii) can be trumped when the information carried by the background beliefs to be given up is too valuable to surrender. Depending upon how valuable that information is, option (i) or (ii) is favored.

The expansion step used to define replacement of $\mathbf{H}^c$ by $\mathbf{H}$ in $\mathbf{K_I}$ is uniquely determined by the contraction of $\mathbf{K_I}$ by removing $\mathbf{H}$. So both options (ii) and (iii) are determined once we can identify the contraction of $\mathbf{K}_I$ by removing $\mathbf{H}^c$. But such specification requires a choice from a roster of contraction strategies each of which removes $\mathbf{H}^c$ from $\mathbf{K}_I$.

Keep in mind that such a "choice" is not a deliberate decision to implement such a contraction. It is not a deliberate choice of a contraction from $\mathbf{K}_\perp$. Nor is it a deliberate choice of a contraction from $\mathbf{K}_I$. If there is any deliberate decision involved it is a deliberate choice of a precommitment at $\mathbf{K_{PI}}$ to a clause in the program for routine expansion precommitting the inquirer to an approach to contracting from inconsistency should routine expansion lead to inconsistency..

Even so, the decision concerning a contingency plan for contracting from inconsistency depends on answering the question: What should be a best or an admissible contraction removing $\mathbf{H}$ from $K_\mathbf{I}$ on the belief contravening supposition that a deliberate choice of a contraction removing $\mathbf{H}^c$ from $\mathbf{K}_I$ is required? The supposition of this question contravenes or should contravene the full beliefs of the inquiring agent. Even so, the question could be addressed as a problem of rational choice. Given a specification of the options and the payoffs, how should a decision be taken? In Levi 1991, I contended that that in such contexts of "coerced" contraction where the inquirer has no coherent option but to retreat from inconsistency, the inquirer still needs to decide "how to contract".

Deliberate Contraction

A decision concerning how to contract also arises in the context of deliberate contraction from consistent $\mathbf{K}_I$ in order to give a hearing to an informationally valuable conjecture.[11]

[11] In Levi 1991, ch. 4, I distinguished between coerced and uncoerced contraction and took note of the fact that both types of contraction require consideration of the problem of "How to Contract".

Sometimes inquirers can be justified in contracting from initial $\mathbf{K}_I$ even though $\mathbf{K}_I$ is consistent. $\mathbf{K}_I$ may be incompatible with $\mathbf{H}$ so that X is committed to being certain that $\mathbf{H}$ is false. Yet, X may regard $\mathbf{H}$ as a hypothesis that would, if it were true, explain propositions in some interesting and important domain.

Thus, the general theory of relativity was incompatible with received doctrine even though it could if true explain the perihelion of Mercury. The behavior of Mercury may have been consistent with $\mathbf{K}_I$ –i.e., with classical mechanics. However, the attempts to explain the perihelion of Mercury within the framework of $\mathbf{K}_I$ were not successful. The trajectory of Mercury represented a stubborn anomaly within the framework of classical mechanics.

Contracting $\mathbf{K_I}$ by removing $\mathbf{H}^c$ so as to recognize the general theory of relativity as a serious possibility is a retreat from anomaly but not from inconsistency. The contraction is deliberately chosen over the option of remaining with K_I in order to give a hearing for an important conjecture even though doing so incurs a loss of informational value.

The loss of informational value might be endured provided the inquirer has confidence that subsequent to contraction, inquiry to resolve the doubt (for example concerning the status of General Relativity) will be undertaken and will lead to removing doubts either by justifying return to $\mathbf{K}_I$ by rejecting General Relativity or expanding $\mathbf{K}_N$ by adding General Relativity. On that assumption, the inquirer can reason that either the informational value will be greater than that carried by $\mathbf{K}_I$ or there will be no net loss in informational value.

This reasoning assumes (1) that the informational value obtained by adding General Relativity Theory to the contraction of $\mathbf{K}_I$ by removing classical gravitational theory is greater than that $\mathbf{K}_I$ and (2) that contraction removing classical gravitational theory from $\mathbf{K}_I$ initiate inquiry warranting a resolution of the doubt. Otherwise reversion to $\mathbf{K}_I$ is justified.

There is another assumption that deserves mention here. It concerns the import of the thesis that the goal of inquiry seeking to modify full beliefs is to increase informational value while reducing risk of error.

According to secular realism the concern to avoid error or reduce risk of error is restricted to the change in belief state being contemplated. It does not concern changes subsequent to the next one. Nor does it concern convergence to the truth in the limit.

The quest for valuable information by way of contrast takes a longer view. In contraction we seek to minimize loss of informational value. But this concern can be trumped by the expectation of a gain in informational value afforded by the subsequent inquiry.

This is the question of determining what should constitute contraction removing h from $\mathbf{K}$. I subsequently discussed this matter in Levi 1997. My final word on this topic (I think) is to be found in Levi 2004. I do not discuss this topic here but place emphasis on what I take to be the two important types of contraction.

There is no inconsistency in endorsing this asymmetry between the concern to avoid error and the aim of maximizing informational value. To be sure, I cannot demonstrate that the goals of inquiry should exhibit the asymmetry. There are many views of the aims of inquiry alternative to the one I am advocating. I think that the view of these goals being proposed gives us a sensible rationale for legitimate routine expansion, justified inductive expansion, coerced contraction and deliberate contraction superior to rival approaches that may be entertained.

Coerced contraction calls for a justification showing how to contract where it is determined already that some contraction from $\mathbf{K}_\perp$ is to be implemented, that the contraction to be implemented is a contraction removing some specific proposition **H** from $\mathbf{K_I}$, but where it is unsettled as to which contraction meeting this requirement to implement as a contraction from $\mathbf{K}_\perp$.

In deliberate contraction, there is no issue concerning retreating from inconsistency. There is a question as to whether to contract by removing proposition **H**, proposition **G**, etc. from $\mathbf{K}_\mathrm{I}$ or not. And given that **H** is to be removed from $\mathbf{K}_\mathrm{I}$, there is a question as to which contraction removing **H** is to be implemented. The last issue raises a problem similar to the problem of how to contract that arises in connection to coerced contraction. Indeed, the similarity between the two problems concerning how to contract entitles us to explore them together.

How to Contract: The Available Options

In any decision problem, the rationality of the choice made depends on the set of options available to the decision maker – that is to say available to the decision maker *according to the decision maker's point of view.* In order for a proposition to represent an available option, the decision maker must be convinced of his or her ability to implement the option if he or she chooses to do so. The decision maker must also judge it a serious possibility that he or she will implement the option. If the decision maker fully believes that he or she will not make the choice, from the decision maker's point of view, deliberation directed at deciding whether to implement the option is pointless.

Deciding how to contract by removing **H** is, so I assume, a problem for choice where the inquirer seeks to implement the best option available that if implemented would remove **H** from $\mathbf{K}_\mathrm{I}$.

Given the nature of problem, the domain of options from which choice should be made should be some subset of the contractions removing **H** from $\mathbf{K_I}$. Should the set of options include all contractions removing **H** from $\mathbf{K_I}$?

This question is not well formed. A contraction removing **H** from $\mathbf{K}_\mathrm{I}$ is a potential state of full belief. We need to identify the Boolean algebra Ω of potential states of full belief and then ascertain the subset of Ω consisting of contractions from $\mathbf{K}_\mathrm{I}$ removing **H**. A customary view is to think of this algebra as the powerset of a set W of atoms conceptually accessible to the inquirer.

There are many reasons for resisting this idea:

(a) Given any set W of atoms, an inquiring agent is conceptually capable of refining elements of this set. This yields a new set of atoms whose powerset consists of elements conceptually accessible to the inquirer. There is no set W of categorically atomic potential states of full belief. There is no set of maximally consistent belief states. There are no possible worlds.
(b) Suppose, however, that we consider a very large domain W and the powerset generated by it. The inquirer may be interested in removing the proposition that G.W.Bush was properly elected President of the USA in 2000 from his or her state of full belief **K**. A potential contraction removing this item from K would be the join of **K** with a set T belonging to the powerset of W each element of which is inconsistent with **K** and at least one of which entails that Bush was not properly elected. But the elements of T might all specify information concerning famines in Africa, recessions in the USA and, indeed, detailed history of the world. This kind of information may be irrelevant according to the inquirer to topic of the inquirer's investigation. What we should be considering is a Boolean algebra of potential states of full belief that the inquirer is committed to judging relevant to this topic.

Instead of beginning with W and its powerset, consider a set U_{LK} of potential states of full belief where a potential state of full belief **LK** entails the truth of exactly one element of U_{LK} and all elements of U_{LK} are consistent with **LK**. Elements of U_{LK} represent the maximally specific potential states of belief that are deemed relevant by the inquirer to the topic under investigation. They generate an algebra of potential states of full belief and they should be conceptually accessible to the inquirer. However, members of U_{LK} need not constitute a logically exhaustive set so that **LK** need not be a logical truth or conceptual necessity. Similarly the members of U_{LK} may be refined if more specific judgments come to be recognized as relevant. The elements of U_{LK} serve as atoms in the Boolean algebra constituted by the powerset of U_{LK}. Elements of the algebra constitute the set of potential states of full belief to be used in trajectories formed by belief changes of relevance in the kind of inquiries under consideration. Perhaps some good reason will arise for refining U_{LK} or adopting a weaker minimal state of full belief. The domain of potential states extends well beyond any atomic algebra we may utilize in characterizing changes in states of full belief.

LK is the *minimal state of full belief* and U_{LK} is the *basic partition* for the purpose of characterizing changes in belief state that are relevant in the context of inquiry into a given budget of problems. The demands for information that are characterized by **LK** and U_{LK} may be modified in the ongoing process of inquiry and different inquiries and their inquirers have different demands that may require the use of different minimal states and basic partititions.

If the inquirer X is in state of full belief **K**, elements of U_{LK} consistent with **K** are elements of the *ultimate partition* U_{K} relative to **K**. Elements of U_{LK} inconsistent with **K** constitute the *dual ultimate partition* U^*_{K}.

Given **LK** and U_{LK}, a potential contraction of **K** is representable by the join of **K** with the join of a subset T of U^*_{K}.[12] The contraction is a contraction removing **H** from **K** if and only if at least one element of T entails $\mathbf{H}^{\mathrm{c}}$.

A maxichoice contraction removing **H** from **K** is a contraction removing **H** from **K** where T contains only one element and it entails $\mathbf{H}^{\mathrm{c}}$.

A contraction removing **H** from **K** is saturatable if and only if exactly one element of T entails $\mathbf{H}^{\mathrm{c}}$,

A partial meet contraction removing **H** from **K** is a contraction removing **H** from **K** where every element of T is a maxichoice contraction entailing $\mathbf{H}^{\mathrm{c}}$.

The options available for choice to the agent X concerned to contract by removing **H** from **K** given **LK** and U_{LK} consist of all contractions removing **H** from **K** relative to **LK** and U_{LK} (or relative to **K** and U^*_{K}).

How to Contract: Minimizing Loss of Informational Value

According to a decision theoretic approach to contraction, one should begin with an explanation of how the goals of contraction determine a value structure for the set of available options (the potential contractions removing **H** from **K**). A value structure (Levi, 1986) is representable by a set of permissible utility functions. An option is V-admissible if and only if it comes out best among the available options according to at least one permissible utility function. In the special case where, the set of permissible utility functions is unique up to a positive affine transformation, the V-admissible options coincide with the set of options that come out best according to all permissible utility functions.

I assume that when the inquirer is to contract in a manner that removes **H** from **K**, he should choose a contraction removing **H** from **K** that minimizes loss of informational value if such an option is available. Every contraction removing **H** from **K** yields a potential state relative to **LK** and U_{LK} that is the join of **K** with the join of a subset T of U_{LK} that contains at least one element that implies $\mathbf{H}^{\mathrm{c}}$. Two factors determine the loss of informational value incurred:

(a) An evaluation of the informational values of elements of U_{LK}. I adopt a numerical measure $cont(\mathrm{x}) = 1 - M(\mathrm{x})$ as the measure of informational value of an element of U_{LK}. Here $M(\mathrm{x})$ is a finitely additive probability measure over the algebra generated by U_{LK}.
(b) An extension of the evaluation of elements of U_{LK} to the entire algebra generated by U_{LK}.

[12]The contraction is also characterized as the intersection of the set of consequences of **K** and the set of consequences of T.

If the extension is the finitely additive probability mentioned in (a), the assessment of informational value is given b y $1 - M(x)$ for all x in the algebra. This is the *undamped* assessment of informational value.

There is an alternative assessment – the *damped assessment of informational value*. Given any finite subset S of the powerset of U_{LK}, the damped informational value of the join of S is the minimum value assigned to an element of S.[13]

In discussing contraction, we are interested in subsets of the dual ultimate partition U^*_K given **K**. The dual ultimate partition is, of course, a subset of the basic partition. The informational value of a subset T of U^*_K is the minimum of the informational values assigned elements of T and hence the maximum value of informational value determining probability M. This maximum exists as long as T is finite. We shall consider cases where U^*_K and, hence, T is finite.

In Levi 2004, I proposed that contraction from **K** removing **H** should be assessed by evaluating the damped informational value of every contraction from **K** removing **H** recognized by **K** and U^*_K and restricting choice to contractions that minimize loss of damped informational value.

I also proposed a rule for Ties that recommended choosing the weakest contraction minimizing loss of informational value.

This proposal is intended for cases where the inquirer is committed to evaluating damped informational value based on a numerically determinate probability distribution over U_{LK}.

In general, inquirers will not be committed to such numerically determinate assessments but rather to a set of permissible assessments. I contend that the set should be convex. The recommendation is to choose the weakest of the *V-admissible* contraction.

The result is that the appropriate contraction removing **H** from **K** should be what Rott and Pagnucco call "severe withdrawal" and I call "mild contraction" for situations where the evaluation of loss of informational value is numerically determinate or is representable the restriction of the set of probability distributions over U_{LK} to elements of U^*_K. that weakly order the elements of U^*_K and, hence, maxichoice contractions from **K** whether they remove **H** or not.

Moreover, the recommendation is based on appeal to an account of rational choice appropriate both for cases where the options are weakly ordered and where weal ordering fails.

If the domain of options were to be restricted to the domain of partial meet contractions removing **H** from **K**, and the assessment of informational remains numerically determinate or at least free from ordinal conflict, the recommendation would coincide with the recommendation of those who subscribe to the AGM account of contraction.

[13]In Levi 1991 and 1997, I proposed a variant on the damped informational value assessment. I abandoned it in Levi 2004. I called the earlier variant "version 1 damped informational value" and the later one version 2 damped informational value. In Rott (2006), Hans Rott showed that my characterization of version 1 was seriously defective. In Levi (2006), I showed how to repair the defect. But there were other reasons including comments by Hansson and Olsson (1995) that argue for abandoning version 1 damped informational value. I rehearse them in Levi (2004).

But the AGM account provides no decision theoretic basis for restricting the domain of options to just the partial meet contractions.

I do not mean to suggest that arguments have not been given in support of partial meet contraction. David Makinson has offered at least two. The first points out that every contraction removing **H** from **K** that is not the join of a subset T of U^*_K all of whose members entail $\mathbf{H}^c$ is the join of a subset T' that is the union of just such a subset with another whose members entail **H**. This "withdrawal incurs a greater loss of information than is incurred by using the partial meet contraction.

If we think of this argument decision theoretically, it presupposes that the goal of contraction should be to minimize loss of information rather than informational value. This is not a goal to which many, including Makinson, would subscribe; for if endorsed it favors restricting choice to maxichoice contractions removing **H** from **K**.[14]

[14]H. Rott (2006) declares that my recent proposal for assessing loss of informational value (damped informational value version 2) is strongly counterintuitive. The basis for his charge is that the recommended contraction could be the disjunction of a great number of maxichoice contractions each of which on its own incurs a great loss of informational value. But this basis presupposes what I deny – to wit, that the aim of contraction is to minimize loss of undamped informational value. Rott also insists that my proposal is sheer definition or stipulation. The only sense in which this is so is that damped informational value version 2 represents one among many utility functions that might be candidates for representing the value of information. I do not deny that rational agents could have different preferences. But the alternatives do not appear appetizing. Consider the interesting results reported in Rott (1993, 2001). Rott has explored choice functions over a domain of maxichoice functions and the preference relations thus generated. In cases where the choice function yields a set of two or more maxichoice contractions as optimal, he *stipulates without appeal to decision theoretic considerations* implementation of a corresponding partial meet contraction. Having done this he establishes a connection between choice consistency conditions for choice functions defined over the domain of maxichoice contractions and important axioms for contraction. Had he adopted a version 2 damped informational value utility function over the power set of the maxichoice contractions removing **H** the correspondence Rott identifies could have been rationalized decision theoretically rather than by stipulation. Even with this improvement, the approach still restricts the options to the set of partial meet contractions without any decision theoretic rationale for doing so whatsoever. I do not deny that Rott has established a correspondence between choice consistency over maxichoice contractions and axioms for contraction. Such a correspondence may satisfy philosophical logicians. But I do not see how it could satisfy anyone interested in a decision theoretic rationalization of contraction.

Pagnucco and Rott (1999) do consider the full range of contractions removing **H** as options. But, as Rott says, they think of the goals of contraction in such cases as being informational value and fairness (expressed by rule for Ties) which are conflicting primary desiderata.

I cannot prove anyone irrational who takes this position but I think it is obviously untenable. When two potential contractions removing **H** minimize loss of informational value, one may break ties by suspending judgment – which involves moving to a weaker contraction than either of the two – *provided that this contraction does not incur a greater loss in informational value*. If it does, invoking the rule for Ties is untenable. If it does not, then even if the rule for ties recommends favoring a contraction weaker than the two options that carry the same informational value that it does, it recommends minimizing loss of informational value. Why should Pagnucco and Rott deny this? I suspect that they think that suspending judgment between two options that minimize loss of information or informational value cannot coherently minimize loss of information or of informational value. This is so for information. But why do they insist that it is so for informational

The second argument alludes to the fact that every withdrawal is revision equivalent to a partial meet contraction. That is to say, expanding the partial meet contraction and expanding the revision equivalent withdrawal yield the same revision of K. Here the revision is an AGM revision.

Consider the following transformation of **K** by adding **K**. The Ramsey revision of **K** by adding **H** is the same as AGM revision in two cases: (i) **K** is inconsistent with **H**, (ii) **K** is consistent with both **H** and $\mathbf{H}^c$. In case (iii) where **K** is inconsistent with $\mathbf{H}^c$, AGM revision by adding **H** is an identity transformation. Ramsey revision requires contraction by removing **H** and then expanding by adding **H**.

If the Recovery postulate for contraction is applicable, AGM revision and Ramsey revision are equivalent. But Recovery should fail in many cases – most notably statistical examples. If X knows that coin has been tossed and landed heads and then contracts by removing the claim that the coin has been tossed, X should give up the claim that the coin landed heads (or, indeed, that it landed at all). Restoring the claim that the coin has been tossed will restore the claim that the coin landed on the surface but not that it landed heads. In that case recovery fails and AGM revision is no longer equivalent to Ramsey revision.

Although every withdrawal is AGM revision equivalent to a partial meet contraction, it is not Ramsey revision equivalent to a partial meet contraction. If the view of legitimate belief change I have been sketching is along right lines, neither form of revision plays a central role in justifying belief change. Revision comes into its own when an analysis of modal judgment on a supposition is on offer. I contend that such Ramsey revision is more adequate to this task than AGM revision.

Hans Rott (1993, 2001) has offered a representation of the preference among maxichoice contractions in terms of choice functions. Given a choice function over the domain of maxichoice contractions removing **H**, the value of the function is the set of optimal maxichoice contractions. removing **H**. The contraction determined by the choice function is the join of the set of optimal maxichoice contractions and is the partial meet contraction. This join is not, in general, a maxichoice contraction and, hence, is not a member of the value of the choice function examined by Rott. But it is join of the maxichoice contractions in the value of the choice function.

One could examine choice functions that take as arguments sets of contractions removing **H** whether they are maxichoice or not. Rott does not develop an account of preference over all contractions that shows that the meet contraction is best among all the options available to the decision maker. He does show that the most preferred maxichoice contractions removing **H** from **K** can be used to define a partial meet contraction. But given his restricted account of the domain of the choice functions, Rott cannot show that the recommended partial meet contraction is optimal.

It would have been easy for Rott to have obtained this definition of partial meet contraction in terms of choice functions (at least in the finite case) had he extended his choice functions from the domain of maxichoice contractions removing **H** from

value? If it were true, minimizing loss of informational value would lead straightforwardly to choosing maxichoice contractions a consequence he acknowledges is untenable.

K to the powerset of this domain (that is to say, the set of partial meet contractions). Given any set of maxichoice contractions, its value would be the best of the values of the maxichoice contractions.

But even if Rott had taken this step, he would not have provided a satisfactory decision theoretic account of contraction because he would not have explained decision theoretically why "withdrawals" are left out of account. Using damped informational value (that is to say the type 2 or second version), it is possible to account for withdrawals.

My aim here has been to sketch the rationalization I have offered for a decision theoretic approach to how to contract that completes the account of coerced contraction and deliberate contraction outlined in previous sections. Given that account, the corrigibility of the inquirer's point of view, from the inquirer's point of view, is justifiable even though the inquirer to be coherent is committed to ruling out the serious possibility that his or her current point of view is in error.

Curiosity and Corrigibilism

As I have done throughout most of my career, I have been maintaining that the belief states that are targets for justifiable change are states of full belief. Such states serve as standards for serious possibility. From the point of view of the inquirer who makes the judgment of serious possibility, there is no serious possibility that what the inquirer fully beliefs is false. The inquirer is committed to epistemological infallibilism.

The main philosophical tradition maintains that such a view is untenable because it implies epistemological incorrigibilism. I deny this. The inquiring agent is sometimes warranted in contracting his or her state of full belief.

My argument for this view is based on the legitimacy of both routine expansion and deliberate or inductive expansion.

Routine expansion is potentially conflict injecting. If we are to acknowledge the legitimacy of such expansion whether by appealing to the testimony of the senses or of competent witnesses and experts, programs for routine expansion must provide contingency plans for contraction in case inconsistency inadvertently arises.

Deliberate contraction to give informationally valuable propositions a hearing can be rationalized only on the assumption that subsequent to such contraction, expansion will be legitimate that affords a promise of removing the doubts that are raised. Sometimes routine expansion may be all that is required. But when the conjectures to be given a hearing are highly theoretical (as is the case with General Relativity Theory), deliberate or inductive expansion may be needed.

Expansion, so I claim, is legitimate as long as the quest for valuable information may be seen to compensate for the risk of importing error incurred in expansion. If we are to avoid the dogma that equates infallilibilism with incorrigibilism, we need to follow William James in rejecting W.K. Clifford's emphasis on the avoidance of error as the sole desideratum of inquiry.

References

Alchourrón, C., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet functions for contraction and revision. *Journal for Symbolic Logic, 50*, 510–530.

Carnap, R. (1960). The aim of inductive logic. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science* (pp. 302–318). Stanford: Stanford University Press.

Cohen, L. J. (1970). *The implications of induction*. London: Methuen.

Cohen, L. J. (1977). *The probable and the provable.* Oxford: Oxford University Press.

Davidson, D. (1998). Truth rehabilitated, Unpublished Manuscript.

De Finetti, B. (1964). Foresight: Its logical laws, its subjective sources. In H. E. Kyburg & H. Smokler (Eds.), *Studies in subjective probability* (pp. 93–158). New York: Wiley.

Hansson, S. O., & Olsson, E. J. (1995). Levi contractions and AGM contractions: a comparison. *Notre Dame Journal of Formal Logic, 36*, 103–119.

Jeffrey, R. C. (1965). *The logic of decision*. New York: McGraw Hill.

Levi, I. (1967a). *Gambling with truth*. Cambridge: MIT Press, Paper, 1973.

Levi, I. (1967b). Information and inference. *Synthese, 17*, 369–391.

Levi, I. (1974). On indeterminate probabilities. *The Journal of Philosophy, 71*, 391–418.

Levi, I. (1980). *The enterprise of knowledge*. Cambridge: MIT Press, Paper, 1983.

Levi, I. (1983). Truth, fallibility and the growth of knowledge. In R. S. Cohen & M. W. Wartofsky (Eds.), *Language, logic and method* (pp. 153–174). Dordrecht: Reidel.

Levi, I. (1986). *Hard choices: Decision making under unresolved conflict*. New York: Cambridge University Press.

Levi, I. (1991). *The fixation of belief and its undoing*. Cambridge: Cambridge University Press, Paper, 2009.

Levi, I. (1997). *For the sake of the argument*. Cambridge: Cambridge University Press, Paper, 2007.

Levi, I. (2002). Maximizing and satisficing measures of evidential support. In M. David (Ed.), *Reading natural philosophy: Essays in the history and philosophy of science and mathematics* (pp. 315–333). Chicago: Open Court.

Levi, I. (2003). Contraction from epistemic hell is routine. *Synthese, 135*, 141–164.

Levi, I. (2004). *Mild contraction*. Oxford: Oxford University Press.

Levi, I. (2006). Informational value should be relevant and damped! Reply to Rott (2006).

Niiniluoto, I. (1984). *Is science progressive?* Dordrecht: D. Reidel.

Olsson, E. J. (2003). Avoiding epistemic hell, Levi on testimony and consistency. *Synthese, 135*, 119–140.

Pagnucco, M., & Rott, I. (1999). Severe withdrawal and recovery. *Journal of Philosophical Logic, 28*, 501–547. See 'Erratum' *Journal of Philosophical Logic* 29 (2000).

Ramsey, F. P., (1990). *Philosophial papers* (D. H. Mellor, Ed.). Cambridge: MIT Press.

Rott, H. (1993). Belief contraction in the context of the general theory of rational choice. *Journal of Symbolic Logic, 58*, 1426–1450.

Rott, H. (2001). *Change, choice and inference: A study of belief revision and nonmonotonic reasoning*. Oxford: Oxford University Press.

Rott, H. (2006). The value of truth and the value of information: On Isaac Levi's epistemology. In E. Olsson (Ed.), *Knowledge and inquiry* (pp. 179–200). Cambridge: Cambridge University Press.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Shackle, G. L. S. (1949, 1952). *Expectation in economics,* Cambridge: Cambridge University Press.

Shackle, G. L. S. (1961, 1969). *Decision, order and time*. Cambridge: Cambridge University Press.

Spohn, W. (1988). Ordinal conditional functions: A dynamic theory of epistemic states. In W. Harper & B. Skyrms (Eds.), *Causation in decision, belief change and statistics* (pp. 105–134). Dordrecht: Kluwer Academic Publishers.

Chapter 16
Belief Contraction in the Context of the General Theory of Rational Choice

Hans Rott

Introduction

The theory of partial meet contraction and revision was developed by Alchourrón, Gärdenfors and Makinson (henceforth, "AGM") in a paper published in the *Journal of Symbolic Logic* in 1985. That paper is the by now classic reference of the flourishing research programme of *theory change*, or as it is alternatively called, *belief revision* (Fuhrmann and Morreau 1991; Gärdenfors 1988, 1992; Katsuno and Mendelzon 1991). In particular, it has been shown that partial meet contraction is a powerful tool for the reconstruction of other kinds of theory change such as safe contraction, epistemic entrenchment contraction, and base contraction (Alchourrón and Makinson 1986; Nebel 1989, 1992; Rott 1991, 1992a).

The basic idea of partial meet contraction is as follows. In order to eliminate a proposition x from a theory A while obeying the constraint of deductive closure and minimizing the loss of information, it is plausible to look at the maximal subsets B of A that fail to imply x. In an earlier paper, Alchourrón and Makinson had proved that when $A = Cn(A)$ then taking one such B leaves an agent with too many propositions, while taking the intersection of all such B's leaves him with too few. In AGM (1985), AGM investigate the idea of taking the intersection of *a select set* of such B's. The choice which B's to take is made with the help of a *selection function*. A natural question is whether all these selections can be represented as the selections of *preferred* B's, where the preferences between maximally nonimplying subsets of A are independent of the proposition x to be deleted.

The purpose of the present paper is threefold. First, we put the theory of partial meet contraction in a broader perspective. We decompose it into two layers, each

H. Rott (✉)
Department of Philosophy, University of Regensburg, 93040 Regensburg, Germany
e-mail: hans.rott@ur.de

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_16

of which can be cultivated with the help of methods developed in different research areas. On the one hand, in relating selection functions to preference relations we can draw on existing work in social choice theory (Chernoff 1954; Herzberger 1973; Sen 1982; Suzumura 1983). On the other hand, we shall elaborate on a remark of Grove (1988) and link maximally nonimplying subsets to "possible worlds" or models, thereby making it possible to compare partial meet contraction with semantical approaches to belief revision and nonmonotonic reasoning (Kraus et al. 1990; Katsuno and Mendelzon 1991; Lindström 1991; Lehmann and Magidor 1992). Exaggerating somewhat, we can say that the theory of partial meet contraction emerges from juxtaposing the general theory of rational choice and a tiny bit of model theory. After introducing abstract postulates for contraction functions, we reprove the two main representation theorems of AGM (1985, pp. 521, 530) concerning partial meet contraction and transitively relational partial meet contraction in a slightly more systematic fashion.

Second, we provide a partial solution to a problem left unanswered by AGM and still considered to be an interesting open question in Makinson and Gärdenfors (1991, p. 195). More precisely, we present two new results that lie strictly "between" those of AGM (1985), viz., representation theorems for *relational* and *negatively transitively relational* partial meet contraction. However, these results hold only under certain preconditions. If the theory to be contracted is logically finite, then all these conditions are met. Our decomposition allows for a uniform method of proof using so-called Samuelson preferences. It increases our understanding of the partial meet contraction mechanism by localizing exactly in which parts of the proofs the finiteness assumption is effective.

Third, as an application, we explore the logic of a variant of syntax-based belief change, namely *simple* and *prioritized base contractions* in the finite case. The basic idea here is that real life theories are generated from finite axiomatizations and that the axioms may carry different epistemic weight. Both the syntactical encoding of a theory and the prioritization are held to be relevant for theory change. We achieve a logical characterization of simple and prioritized base contraction by combining one of our representation theorems with observations originating with Lewis (1981) and Nebel (1989, 1992).

Independently from the research for this article, related work has been carried out by Katsuno and Mendelzon (1991) and Lindström (1991). Lindström proves numerous representation theorems, and like us, he heavily draws on results from the general theory of rational choice. However, there are some important differences. He adopts a semantic approach right from the outset, while our reconstruction starts by turning a syntactic approach into an essentially semantic one, with the notable feature that any two "worlds" in our sense are linguistically distinguishable (our models are "injective", to use a term of Michael Freund). Lindström's main concern is nonmonotonic inference relations and the related area of belief *revision*, whereas we shall focus on belief *contraction*. (Everything we might wish to say about the intimate connection between revisions and contractions is contained in AGM 1985 and Gärdenfors 1988). He applies *liberal* maximization where we apply *stringent* maximization. He uses different postulates for choice functions and a

different notion of revealed preference. The main contrast between his approach and ours, however, is that Lindström permits revisions by—possibly infinite—*sets* of propositions whilst we stick to the original AGM model of changing theories by *single* propositions. Lindström's generalization allows him to revise by "worlds" and thus dispense with a finiteness assumption which will prove to be necessary at several places in the AGM framework. In fact, it is a major aim of our paper to make transparent the reasons where and why logical finiteness is needed in the AGM theory. Further important results on purely finitistic belief revision are found in Katsuno and Mendelzon (1991) who incorporate a revision operator into their object language. Both papers are highly recommended and should be read in conjunction with the present one.

Unfortunately, space limitations do not allow a presentation that makes for easy reading. Familiarity with earlier work either in the area of theory change—an excellent overview is Gärdenfors (1988)—or in the general theory of rational choice—of particular relevance is Herzberger (1973)— will greatly facilitate the reader's task. However, we will shortly repeat the basic definitions so that our presentation is in principle self-contained. It may be useful to inform the reader in advance about the fundamental entities that will show up as determinants of theory change. We shall meet contraction functions $\div$, selection functions γ, preference relations $\leq$ (and $\underline{\leq}$), maximally nonimplying sets M (and N), maximally consistent sets or "worlds" W, as well as simple and prioritized belief bases B and $\langle B_i \rangle$. We shall frequently jump to and fro between these kinds of entities. When we wish to generate a *c*ontraction function, *s*election function, *p*reference relation, *m*aximal nonimplying set, *w*orld, or belief *b*ase from another kind of entity, we shall use metalevel mappings denoted by $\mathscr{C}, \mathscr{S}, \mathscr{P}, \mathscr{M}, \mathscr{W}, \mathscr{B}$ respectively.

Selection Functions and Preference Relations

General Selection Functions

Let X be a set and $\mathscr{X}$ be a nonempty subset of $2^X - \{\emptyset\}$. A *selection function* (or *choice function*) over $\mathscr{X}$ is a function $\gamma : \mathscr{X} \to 2^X$ such that $\gamma(S)$ is a nonempty subset of S, for every $S \in \mathscr{X}$. Intuitively, selection functions are supposed to give us the best elements of each S in $\mathscr{X}$. The requirement that $\gamma(S)$ be nonempty means that a selection function is effective or "decisive": In every case it reaches a decision of which elements are best. As Lindström (1991) points out, the decisiveness of selection functions corresponds to a condition of consistency preservation in belief revision and nonmonotonic reasoning. Here consistency is judged by *some* monotonic background consequence relation, not necessarily by the classical one.

Domain conditions for selection functions are important. A set $\mathscr{X}$ of subsets of X is called *n-covering* ($n = 1, 2, 3, \ldots$) if it contains all subsets of X with exactly n elements, $\mathscr{X}$ is called *$n_1 n_2$-covering* (or *$n_1 n_2 n_3$-covering*) if it is n_1-covering and

n_2-covering (and n_3-covering). $\mathscr{X}$ is called *ω-covering* if it is n-covering for all natural numbers $n = 1, 2, 3, \ldots$. The set $\mathscr{X}$ is called *additive* if it is closed under arbitrary unions, and it is called *finitely additive* if it is closed under finite unions; it is called *subtractive* if for every S and S' in $\mathscr{X}$ such that $S \not\subseteq S'$, $S - S'$ is also in $\mathscr{X}$, and it is called *full* if $\mathscr{X} = 2^X - \{\emptyset\}$. Of course, if $\mathscr{X}$ is 1-covering and finitely additive, then it is ω-covering. Finally, we say that $\mathscr{X}$ is *compact* if for every T and S_i, $i \in I$, from $\mathscr{X}$, if $T \subseteq \bigcup\{S_i : i \in I\}$ then $T \subseteq \bigcup\{S_i : i \in I_0\}$ for some finite $I_0 \subseteq I$. We say that a *selection function* γ with domain $\mathscr{X}$ is n-covering, additive, subtractive, etc., if its domain $\mathscr{X}$ is n-covering, additive, subtractive, etc.

Constraints on Selection Functions

We consider the following "coherence postulates" for selection functions. Their motivation, where it is not obvious, is given in Herzberger (1973), Sen (1982), and Suzumura (1983).

(I) For all $S, S' \in \mathscr{X}$, if $S \subseteq S'$ then $S \cap \gamma(S') \subseteq \gamma(S)$ ("Chernoff's axiom", Sen's Property α)

(II) For all $\{S_i : i \in I\} \subseteq \mathscr{X}$ such that $\bigcup\{S_i : i \in I\} \in \mathscr{X}$, $\bigcap\{\gamma(S_i) : i \in I\} \subseteq \gamma(\bigcup\{S_i : i \in I\})$ (Sen's Property γ)

(III) For all $S, S' \in \mathscr{X}$ such that $S \subseteq S'$, if $\gamma(S') \subseteq \gamma(S)$ then $\gamma(S) \subseteq \gamma(S')$ ("Superset Axiom", Sen's Property ε)

(IV) For all $S, S' \in \mathscr{X}$ such that $S \subseteq S'$, if $\gamma(S') \cap S \neq \emptyset$, then $\gamma(S) \subseteq \gamma(S')$ ("Arrow's axiom", Sen's Property $\beta+$)

Condition (III) is independent of condition (II), even for finite X and in the presence of condition (I). Example: Let $X = \{x, y, z\}$, $\mathscr{X}$ be the power set of X minus $\emptyset$, and consider γ_1 and γ_2 defined by $\gamma_1(\{x, y\}) = \gamma_2(\{x, y\}) = \{x, y\}$, $\gamma_1(\{x, z\}) = \{x\}$, $\gamma_1(\{y, z\}) = \{z\}$, $\gamma_1(\{x, y, z\}) = \{x\}$, $\gamma_2(\{x, z\}) = \{x, z\}$, $\gamma_2(\{y, z\}) = \{y, z\}$, and $\gamma_2(\{x, y, z\}) = \{x, y\}$. Evidently, γ_1 satisfies (I) and (II) and violates (III), while γ_2 satisfies (I) and (III) and violates (II). On the other hand, condition (IV) implies condition (II): Let $\{S_i : i \in I\} \subseteq \mathscr{X}$, $\bigcup\{S_i : i \in I\} \in \mathscr{X}$. Now $\emptyset \neq \gamma(\bigcup\{S_i : i \in I\}) \subseteq \bigcup\{S_i : i \in I\}$, so $\gamma(\bigcup\{S_i : i \in I\}) \cap S_{i_0} \neq \emptyset$ for some $i_0 \in I$. So by (IV), $\bigcap\{\gamma(S_i) : i \in I\} \subseteq \gamma(S_{i_0}) \subseteq \gamma(\bigcup\{S_i : i \in I\})$. It is trivial to see that condition (IV) also implies condition (III), since $\emptyset \neq \gamma(S') \subseteq \gamma(S)$ implies $\gamma(S') \cap S \neq \emptyset$.

Rational Choice

In the general theory of choice and preference we often find an idea which can be phrased in the slogan "*Rational choice is relational choice*". That is, rational choice is choice which can be construed as based on an underlying preference relation.

> The intended interpretation of the set $\gamma(S)$, called the *choice set* for S, is that its elements are regarded as equally adequate or satisfactory choices for an agent whose values are represented by the function γ, and who faces a decision problem represented by the set S. Following Chernoff (1954) ..., this relativistic concept of equiadequacy for a given decision problem bears sharp distinction from invariant concepts like preferential matching or indifference which for a given agent are not relativized to decision problems, and which may be subject to more stringent constraints, both for rational agents and for agents in general. (Herzberger 1973, p. 189, notation adapted)

Choice sets are taken to be sets of "best" elements. There are basically two ideas to make this precise. The first is based on a nonstrict (reflexive) preference relation $\leq$:

$$\gamma(S) = \{y \in S : y' \leq y \text{ for all } y' \in S\}$$

The second idea is based on a strict (asymmetric) preference relation $<$:

$$\gamma(S) = \{y \in S : y \not< y' \text{ for all } y' \in S\}$$

These suggestions are respectively referred to as *stringent* and *liberal maximization* in Herzberger (1973, p. 197), and *G-rationality* and *M-rationality* in Suzumura (1983, p. 21). (Stringent maximization is often attributed to Condorcet 1785.) If $<$ is the converse complement of $\leq$, then stringent and liberal maximization coincide. If $<$ is the asymmetric part of $\leq$, then every liberal maximizer with respect to $<$ is a stringent maximizer with respect to the *augmentation* $\leq^+$ of $\leq$ which is defined by $x \leq^+ y$ iff $x \leq y$ or $y \not\leq x$, i.e., iff not $y < x$. Clearly, $\leq^+$ is *connected*, i.e., it holds for every x and y that either $x \leq^+ y$ or $y \leq^+ x$. While $\leq$ allows us to keep apart indifferences (both $x \leq y$ and $y \leq x$) from incomparabilities (neither $x \leq y$ nor $y \leq x$), $\leq^+$ blurs just this distinction, for whenever we have neither $x \leq y$ nor $y \leq x$ we have both $x \leq^+ y$ and $y \leq^+ x$. Thus, we may not expect $\leq^+$ to be transitive. If, on the other hand, $\leq$ is already connected then $\leq^+ = \leq$, and stringent maximization with respect to the asymmetric part $<$ of $\leq$ coincides with liberal maximization with respect to $\leq$. For all this, cf. Herzberger (1973, § 3).

In accordance with AGM (1985) as well as with the dominant approach in the theory of choice and preference, we shall focus on stringent maximization.[1] From now on, when we say that γ is *relational with respect to* $\leq$ *over* X, we mean that $\gamma(S) = \{y \in S : y' \leq y \text{ for all } y' \in S\}$ for every $S \in \mathscr{X}$. In this case we write

[1] Intuitively, however, I think that liberal maximization is preferable. Liberal maximization is based on strict relations which do not allow to distinguish between incomparabilities and indifferences. Nonstrict relations do make this distinction, but stringent maximization tends to require connected relations which often can be had only if incomparabilities are turned into indifferences—i.e., if augmentations are used. The interpretation of nonstrict relations as the converse complements of—more intuitive—strict relations explains the crucial role of *negative* transitivity and *negative* well-foundedness in the following. Also compare the recommendation in Rott (1992b) to regard the nonstrict epistemic entrenchment relation $\leq_E$ of Gärdenfors and Makinson (1988) as the converse complement of a more intuitive strict relation $<_E$.

$\gamma = \mathscr{S}(\leq)$. When we say that γ is *relational*, we mean that there is some relation $\leq$ over X such that $\gamma = \mathscr{S}(\leq)$.

For any nonstrict relation $\leq$, $<$ is to denote the *asymmetric part* of $\leq$, which is defined by $x < y$ iff $x \leq y$ and not $y \leq x$. $\leq$ is called *n-acyclic*, if no n objects $x_1, x_2, \ldots, x_n$ form a cycle under $<$, i.e., if a chain $x_1 < x_2 < \cdots < x_n < x_1$ does not occur. 1-acyclicity is irreflexivity, 2-acyclicity is asymmetry. $\leq$ is called *acyclic*, if it is n-acyclic for every $n = 1, 2, 3, \ldots$. $\leq$ is called *negatively acyclic* if there is no cycle under $\not\leq$, i.e., if never $x_1 \not\leq x_2 \not\leq \cdots \not\leq x_n \not\leq x_1$, and it is called *negatively well-founded* if there is no infinite descending chain under $\not\leq$, i.e., if never $\cdots \not\leq x_3 \not\leq x_2 \not\leq x_1$. For connected relations $\leq$, negative well-foundedness coincides with converse well-foundedness where infinite ascending chains $x_1 < x_2 < x_3 < \ldots$ do not occur. Obviously, if $\leq$ is conversely (or negatively) well-founded then it is acyclic (negatively acyclic). $\leq$ is *smooth* with respect to $\mathscr{X}$ if there are no infinite descending $\not\leq$-chains in S, for every $S \in \mathscr{X}$. Smoothness is a restricted form of negative well-foundedness. $\leq$ is called *negatively transitive* (*virtually connected, modular, ranked*) if $x \not\leq y$ and $y \not\leq z$ implies $x \not\leq z$. It is quickly verified that a connected relation $\leq$ is negatively transitive iff it is *quasi-transitive* in the sense that its asymmetric part $<$ is transitive. Quasi-transitivity is a much disputed requirement in social choice theory (Herzberger 1973; Sen 1982; Suzumura 1983). It should be noted that all transitive relations $\leq$ are both acyclic and quasi-transitive.

Two Kinds of Revealed Preferences

In many contexts one can hope to recover the underlying preferences of an agent from observed choice behavior. The two most commonly used types of "revealed preference" relations are the *Samuelson preferences* (Samuelson (1950))

$$\leq_\gamma = \{\langle x, x'\rangle \in X \times X : \exists S \in \mathscr{X} \text{ such that } \{x, x'\} \subseteq S \quad \text{and } x' \in \gamma(S)\} \tag{16.1}$$

$$= \bigcup \{S \times \gamma(S) : S \in \mathscr{X}\} \tag{16.2}$$

and the so-called *base preferences* (Uzawa (1956); Arrow (1959)):

$$\leq_{\gamma,2} = \{\langle x, x'\rangle \in X \times X : \{x, x'\} \in \mathscr{X} \text{ and } x' \in \gamma(\{x, x'\})\} \tag{16.3}$$

$$= \bigcup \{S \times \gamma(S) : S \in \mathscr{X} \text{ and } S \text{ has at most two elements}\} \tag{16.4}$$

The terminology is taken from Herzberger (1973), where many other possibilities of defining notions of revealed preference are discussed. Obviously, $x \leq_{\gamma,2} y$ implies $x \leq_\gamma y$. Notice that neither of these relations is guaranteed to be reflexive, unless γ is 1-covering. $\leq_{\gamma,2}$ is defined for arbitrary γ's, but the definition makes good sense only for 2-covering ones for which $\{x, x'\}$ is always in $\mathscr{X}$. In this case, base

preferences and Samuelson preferences are connected. Given a selection function γ, we shall also denote $\preceq_\gamma$ by $\mathscr{P}(\gamma)$ and $\preceq_{\gamma,2}$ by $\mathscr{P}_2(\gamma)$.[2]

Some Basic Properties and Reformulations

The following lemmas list a number of important facts which are basically common knowledge in the general theory of rational choice (cf. Herzberger 1973; Sen 1982; Suzumura 1983). For the straightforward proofs, see Appendix 2.

Lemma 1. (a) *If γ is relational then it satisfies (I) and (II).*
(b) *If γ is 12-covering and satisfies (I) and (II), then $\gamma = \mathscr{S}(\mathscr{P}_2(\gamma))$.*
(c) *If γ is 12-covering and satisfies (I), then $\mathscr{P}(\gamma) = \mathscr{P}_2(\gamma)$.*
(d) *If γ is 12-covering and satisfies (I) and (II), then $\gamma = \mathscr{S}(\mathscr{P}(\gamma))$.*
(e) *If γ is 12-covering and relational, then $\gamma = \mathscr{S}(\mathscr{P}(\gamma)) = \mathscr{S}(\mathscr{P}_2(\gamma))$.*
(f) *If γ is additive and satisfies (I) and (II), then $\gamma = \mathscr{S}(\mathscr{P}(\gamma))$.*

Lemma 2. (a) *If γ is 12n-covering and satisfies (I), then $\mathscr{P}(\gamma) = \mathscr{P}_2(\gamma)$ is n-acyclic. If γ is ω-covering and satisfies (I), then $\mathscr{P}(\gamma)$ is acyclic.*
(b) *If γ is 123-covering and satisfies (I) and (III), then $\mathscr{P}(\gamma)$ is negatively transitive.*
(c) *If γ is finitely additive and satisfies (IV), then $\mathscr{P}(\gamma)$ is transitive.*

Lemma 3. (a) *If $\leq$ is smooth with respect to $\mathscr{X}$, then $\mathscr{S}(\leq)$ is a selection function over $\mathscr{X}$ which satisfies (I) and (II).*
(b) *If $\leq$ is negatively transitive and negatively well-founded (or: if $\leq$ is negatively transitive and smooth and $\mathscr{X}$ is subtractive), then $\mathscr{S}(\leq)$ satisfies (III).*
(c) *If $\leq$ is transitive, then $\mathscr{S}(\leq)$ satisfies (IV).*

Now we bring conditions (I)–(IV) into a form which is more suitable for our purposes.

(I′) For all $S, S' \in \mathscr{X}$ such that $S \cup S' \in \mathscr{X}$, $\gamma(S \cup S') \subseteq \gamma(S) \cup \gamma(S')$
(II′) For all $S, S' \in \mathscr{X}$ such that $S \cup S' \in \mathscr{X}$, $\gamma(S) \cap \gamma(S') \subseteq \gamma(S \cup S')$
(III′) For all $S \in \mathscr{X}$ and S' such that $S \cup S' \in \mathscr{X}$, if $\gamma(S \cup S') \cap S' = \emptyset$ then $\gamma(S) \subseteq \gamma(S \cup S')$
(IV′) For all $S \in \mathscr{X}$ and S' such that $S \cup S' \in \mathscr{X}$, if $\gamma(S \cup S') \cap S \neq \emptyset$, then $\gamma(S) \subseteq \gamma(S \cup S')$

[2] It is worth pointing out that the characteristic definition of a relation of epistemic entrenchment (see Gärdenfors and Makinson 1988; Rott 1992b) between propositions from an observed contraction behavior, viz.

$$x \preceq_E y \text{ iff } x \notin A\dot{-}(x \wedge y) \text{ or } y \in A\dot{-}(x \wedge y)$$

can also be interpreted as a base preference (Rott 1992b, p. 61). In that paper it is argued that the instruction "remove $x \wedge y$" should be regarded as an instruction to remove x *or* remove y, where the agent holding theory A has free choice which proposition(s) out of $\{x, y\}$ to remove. [Note added in 2015: This sketch of an idea was turned into a theory in Rott (2003).]

(I&II) For all $S, S_i \in \mathscr{X}, i \in I$, if $S \subseteq \bigcup\{S_i : i \in I\}$ then $S \cap \bigcap\{\gamma(S_i) : i \in I\} \subseteq \gamma(S)$
(I&II′) For all $S, S', S'' \in \mathscr{X}$, if $S \subseteq S' \cup S''$ then $S \cap \gamma(S') \cap \gamma(S'') \subseteq \gamma(S)$

Note that condition (II′) is just a restriction of (II) to index sets with at most two elements, and similarly for (I&II) and (I&II′).

Lemma 4. (a) *If γ is subtractive, then condition (I) is equivalent to (I′).*
(b) *If γ is finitely additive, compact, and satisfies (I) and (II′), then $\gamma = \mathscr{S}(\mathscr{P}(\gamma))$.*
(c) *If γ satisfies (I), then (III) is equivalent to (III′).*
(d) *(IV) is equivalent to (IV′).*
(e) *If γ is additive, then the conjunction of (I) and (II) is equivalent to (I&II).*
(f) *If γ is finitely additive, then the conjunction of (I) and (II′) is equivalent to (I&II′).*

Lemma 4(b) is a substitute for Lemma 1(f) in the absence of infinite additivity. The conditions (I&II) and (I&II′) are generalized forms of the conditions (γ7:∞) and (γ7:2) in AGM (1985). AGM (1985, Observation 4.10) noticed that (I&II′) is the key to relationality in partial meet contraction (because the domain $\mathscr{X}$ encountered there is compact), but as they failed to decompose that condition into the simpler ones, (I) and (II′), it remained undigestible for them. As we shall see, they provided a postulate for contraction functions which corresponds to (I), but nothing in AGM (1985) corresponds to (II′).

The results of the theory of rational choice are quite nice, and we may be optimistic about their applicability to theory change as determined by partial meet contraction s. Still there are at least two problems which should not be underestimated. First, we do not know very much about selection functions which are neither 123-covering nor additive.[3] And second, in choice-and-preference theory we are faced with a bewildering pluralism of revealed preference relations (see Herzberger 1973). It is hard to decide in advance which notion of revealed preference is "the" right notion for a given purpose. It will turn out, happily, that the relation $\mathscr{P}(\gamma)$ defined above, and, to a smaller degree, also the relation $\mathscr{P}_2(\gamma)$ which is equivalent in the finite case are suitable for partial meet contraction s.

[3]This is no problem for Lindström (1991) whose selection functions are always ω-covering. Consequently, Lindström's constructions can always make use of the base preferences $\mathscr{P}_2(\gamma)$.

Selection Functions and Preference Relations in Partial Meet Contraction

The General Case

We presuppose a propositional language L, with the usual connectives $\neg, \wedge, \vee, \rightarrow$ and $\leftrightarrow$. From now on, small roman letters $a, b, c, \ldots, x, y, z, \ldots$ denote propositions of L, and capital roman letters $A, B, A', B', \ldots, M, N, M', N' \ldots, W, W', \ldots$ denote sets of propositions. Notational conventions: Capital A is to denote a theory, B an arbitrary set of propositions (in general, a nontheory), M and N maximally nonimplying subsets (of some A or B, respectively), and W a maximally consistent set of L. All these metavariables may occur with primes or subscripts.

As in AGM (1985), L is assumed to be governed by some reflexive, monotonic and idempotent logic (consequence operation) Cn which is supraclassical and compact and satisfies the deduction theorem: $B \subseteq Cn(B)$; if $B \subseteq B'$ then $Cn(B) \subseteq Cn(B')$; $Cn(Cn(B)) = Cn(B)$; if Cn_0 is classical tautological implication then $Cn_0(B) \subseteq Cn(B)$; if $y \in Cn(B)$ then $y \in Cn(B_0)$ for some finite subset B_0 of B; and finally, $y \in Cn(B \cup \{x\})$ iff $x \rightarrow y \in Cn(B)$. That is, Cn is a classical logic in the sense of Lindström (1991). Note that in the present context, idempotence is equivalent to the cut rule, and the deduction theorem is equivalent to disjunction in the antecedents.

We also write $B \vdash x$ for $x \in Cn(B)$. A is a called a *theory* if $A = Cn(A)$. By $[\![x]\!]$ and $[\![B]\!]$ we denote the set of all "possible worlds"(valuations, models) in which x (every x in B) is true. Here and throughout this paper, we identify a "world" satisfying x or B with a maximally Cn-consistent set of propositions which contains x (respectively, B). The set of all "worlds" is denoted $\mathbf{W}$.

Let $B\perp x = \{M \subseteq B : x \notin Cn(M)$ and $x \in Cn(N)$ for all N with $M \subset N \subseteq B\}$. Note that the elements of $B\perp x$ are theories if B is a theory. If $\nvdash x$, then $B\perp x$ is nonempty, by the compactness of Cn.[4] We define $B\perp = \{B\perp x : x \in Cn(B) - Cn(\emptyset)\}$ and $U_B = \bigcup(B\perp) = \bigcup\{B\perp x : x \in Cn(B) - Cn(\emptyset)\}$. (Here we slightly deviate from the definitions in AGM (1985) which also include x's from $Cn(\emptyset)$.) If $Cn(B) \neq Cn(\emptyset)$, then $B\perp$ is a nonempty subset of $2^{2^B} - \{\emptyset\}$. The case $Cn(B) = Cn(\emptyset)$ will be handled separately in our reconstruction of partial meet contraction.

Lemma 5. *Let A be a theory and $x \in A$. Then $M \in A\perp x$ iff there is a maximally consistent set ("world") W such that $\neg x \in W$ and $M = A \cap W$.*

For the proof, compare § 4 of Grove (1988).

[4] This marks a difference with Lewis (1981) who identifies propositions with sets of extra-linguistic possible worlds and logical consequence with set-theoretic inclusion. Lacking compactness, Lewis has to ponder the impact of a "Limit Assumption" for premise semantics.

Now let A again be a theory. For $M \in U_A$, put $\mathscr{W}(M) = Cn(M \cup \{\neg x\})$ where x is any proposition in $A - M$, and put $\mathscr{M}(W) = W \cap A$, for any maximally consistent set W such that $A \not\subseteq W$. The reader is invited to verify: $\mathscr{W}$ is well-defined(!), $\mathscr{W}$ is a bijection from U_A to $V_A = \{W \in \boldsymbol{W} : A \not\subseteq W\}$, and $\mathscr{M}$ is the converse of $\mathscr{W}$, i.e., $\mathscr{M}(\mathscr{W}(M)) = M$ and $\mathscr{W}(\mathscr{M}(W)) = W$.

Given this representation of the elements of U_A, it is clear that they satisfy the following *fullness* and *primeness conditions*: If $x \in A - M$ and $y \in A$ then $x \to y \in M$, and if $x, y \in A - M$ then $x \vee y \in A - M$ (cf. Observations 6.1 and 6.2 in AGM 1985). U_A is just the set of all maximal proper subtheories of the theory A. Moreover, we immediately get

Corollary 1. *Let A be a theory and x, y, y_i be in $A - Cn(\emptyset)$.*

(i) *For $M \in U_A$, $x \notin M$ iff $M \in A\!\perp\! x$.*
(ii) $A\!\perp\! (x \wedge y) = A\!\perp\! x \cup A\!\perp\! y$.
(iii) $A\!\perp\! (x \vee y) = A\!\perp\! x \cap A\!\perp\! y$.
(iv) $A\!\perp\! (x \vee \neg y) = A\!\perp\! x - A\!\perp\! y$.
(v) *If $A\!\perp\! x \subseteq \bigcup\{A\!\perp\! y_i : i \in I\}$, then $A\!\perp\! x \subseteq \bigcup\{A\!\perp\! y_i : i \in I_0\}$ for some finite $I_0 \subseteq I$.*

Facts (i) and (ii) are contained in AGM (1985, Lemma 2.4 and Lemma 4.1). We see that $A\!\perp$, the special domain $\mathscr{X}$ of the selection functions which will figure in partial meet contraction s, is closed under finite unions, finite intersections, and differences. We give a direct proof of the compactness property (v). Let $A\!\perp\! x \subseteq \bigcup\{A\!\perp\! y_i : i \in I\}$. Then $\{y_i : i \in I\} \vdash x$. For otherwise, by Lindenbaum's Lemma and Lemma 5, there would be an $M \in A\!\perp\! x$ such that $\{y_i : i \in I\} \subseteq M$, so $M \notin A\!\perp\! y_i$ for every i, contradicting our hypothesis. Compactness of Cn gives us $\{y_i : i \in I_0\} \vdash x$ for some finite $I_0 \subseteq I$. Thus, there is, for every $M \in A\!\perp\! x$, an $i \in I_0$ such that $y_i \notin M$. Hence, by (i), there is, for every $M \in A\!\perp\! x$, an $i \in I_0$ such that $M \in A\!\perp\! y_i$. Thus $A\!\perp\! x \subseteq \bigcup\{A\!\perp\! y_i : i \in I_0\}$, as desired.

Now we can introduce selection functions for belief revision. Let A be a theory. A selection function $\gamma : A\!\perp\! \to 2^{2^A}$ is called a *selection function for A*. It follows from Corollary 1 that every selection function γ for A is finitely additive, subtractive and compact.

Let γ be a selection function for A. The *completion* γ^* of γ is defined by $\gamma^*(A\!\perp\! x) = \{M \in A\!\perp\! x : \bigcap \gamma(A\!\perp\! x) \subseteq M\}$, for all $x \in A - Cn(\emptyset)$. Following AGM (1985), we call a selection function γ *complete* if $\gamma = \gamma^*$. If γ is complete, then $M \in \gamma(A\!\perp\! x)$ whenever $\bigcap \gamma(A\!\perp\! x) \subseteq M \in U_A$ (Proof: Since $\{x \to y : y \in A\} \in \bigcap \gamma(A\!\perp\! x)$, $\bigcap \gamma(A\!\perp\! x) \subseteq M$ implies $M \not\vdash x$, so $M \in A\!\perp\! x$ by Corollary 1(i).)

A *contraction function* over a theory A is a function $\dot{-}_A : L \to 2^L$. We write $A\dot{-}x$ for $\dot{-}_A(x)$, and as there will be no danger of confusion, we shall often write just $\dot{-}$ for $\dot{-}_A$. The term '$A\dot{-}x$' should be read as the 'the result of rationally removing x from A'. The idea of a contraction function dictates that it should satisfy at least the postulates $A\dot{-}x \subseteq A$ and $x \notin A\dot{-}x$ (unless $\vdash x$). More postulates will be discussed in section "Postulates for Contraction Functions". Intuitively, $A\dot{-}x$ is the minimal, most economical or rational, change of A needed to discard x.

A function $\dot{-}$ is the *partial meet contraction function over A determined by the selection function* γ *for A* if and only if

$$A\dot{-}x = \begin{cases} \bigcap \gamma(A\perp x) & \text{for every } x \in A - Cn(\emptyset) \\ A & \text{for every } x \notin A \text{ and every } x \in Cn(\emptyset) \end{cases}$$

Given a selection function γ for A, the partial meet contraction it determines will be denoted by $\mathscr{C}(\gamma)$. There is a slight deviation from AGM (1985) in order to avoid the application of γ to $\emptyset$ and to preserve the perfect correspondence between U_A and V_A. The function $\dot{-}$ is called a *partial meet contraction over A* if there is a selection function γ for A such that $\dot{-} = \mathscr{C}(\gamma)$.

A selection function γ for A is called *(transitively, quasi-transitively, connectively, acyclicly) relational over A* if there is a (transitive, quasi-transitive, connected, acyclic) preference relation $\leq$ over U_A (over 2^A in AGM 1985) such that for every $x \in A - Cn(\emptyset)$:

$$\gamma(A\perp x) = \{M \in A\perp x : M' \leq M \text{ for all } M' \in A\perp x\}.$$

This is an application of stringent maximization as discussed in section "Rational Choice", and it is called the *marking off identity* in AGM (1985). Recall that we write $\gamma = \mathscr{S}(\leq)$ if γ is determined by $\leq$ in that way.

A contraction function $\dot{-}$ is *(transitively, quasi-transitively, connectively, acyclicly) relational* if there is a (transitive, quasi-transitive, connective, acyclic) preference relation $\leq$ over U_A such that $\dot{-} = \mathscr{C}(\mathscr{S}(\leq))$.

Using the above functions $\mathscr{W}$ and $\mathscr{M}$, we find that the selection γ of sets in $A\perp x$ can equivalently be viewed as a selection γ_W of maximally consistent sets, or "worlds", in $[\![\neg x]\!]$. If we define

$$\gamma_W([\![\neg x]\!]) = \{\mathscr{W}(M) : M \in \gamma(A\perp x)\}$$

then clearly,

$$\gamma(A\perp x) = \{\mathscr{M}(W) : W \in \gamma_W([\![\neg x]\!])\}$$

In the principal case $x \in A - Cn(\emptyset)$, $A\dot{-}x = A \cap \bigcap(\gamma_W([\![\neg x]\!]))$. Conditions for the partial meet mechanism may thus be viewed as *model theoretic* or *semantical* conditions. The "possible worlds" view facilitates a visualization of the conditions on preference relations and selection functions and relates the work on partial meet contraction directly to the minimal models approach as adopted in Lewis (1973, 1981), Kraus et al. (1990), Katsuno and Mendelzon (1991), Lindström (1991), and Lehmann and Magidor (1992).

A set W' of maximal consistent sets of propositions is called *elementary* (Δ-*elementary*) if there is a proposition y (a set of propositions B) such that W' is exactly the set of all maximally consistent sets of propositions which contain y (respectively,

which include B), in symbols $W' = [\![y]\!]$ ($W' = [\![B]\!]$). We call γ *elementary* (Δ-*elementary*) if $\gamma_W([\![\neg x]\!])$ is elementary (Δ-elementary) for every $x \in A - Cn(\emptyset)$. If γ is elementary (Δ-elementary) and $\gamma_W([\![\neg x]\!]) = [\![y]\!]$ ($\gamma_W([\![\neg x]\!]) = [\![B]\!]$), then $A \div x = Cn(\{z \vee y : z \in A\}) = A \cap Cn(y)$ (respectively, $A \div x = Cn(\{z \vee z' : z \in A$ and $z' \in B\}) = A \cap Cn(B)$).

Remark 1. γ is complete iff it is Δ-elementary.

Proof. From left to right. We show that $[\![(\bigcap \gamma(A\perp x)) \cup \{\neg x\}]\!] = \gamma_W([\![\neg x]\!])$, for every $x \in A - Cn(\emptyset)$. Clearly, $(\bigcap \gamma(A\perp x)) \cup \{\neg x\} \subseteq W$ for every $W \in \gamma_W([\![\neg x]\!])$. To show the converse, suppose for reductio that $(\bigcap \gamma(A\perp x)) \cup \{\neg x\} \subseteq W$ and $W \notin \gamma_W([\![\neg x]\!])$. Hence, by the latter, $\mathscr{M}(W) \notin \gamma(A\perp x)$, so by the completeness of γ, $\bigcap \gamma(A\perp x) \not\subseteq \mathscr{M}(W)$. But also $(\bigcap \gamma(A\perp x)) \cup \{\neg x\} \subseteq W$, and $\bigcap \gamma(A\perp x) \subseteq A$, so $\bigcap \gamma(A\perp x) \subseteq \mathscr{M}(W)$, and we have a contradiction.

From right to left. Suppose for reductio that there are $x \in A - Cn(\emptyset)$ and $M \in U_A$ such that $\bigcap \gamma(A\perp x) \subseteq M$ and $M \notin \gamma(A\perp x)$. Hence $\mathscr{W}(M) \notin \gamma_W([\![\neg x]\!])$. Since γ is Δ-elementary, there is a set B of propositions such that $\gamma_W([\![\neg x]\!]) = [\![B]\!]$. So $B \not\subseteq \mathscr{W}(M)$. Take some $y \in B - \mathscr{W}(M)$. Since $x \vee y \notin \mathscr{W}(M)$, we get, by the definition of $\mathscr{W}$, $x \vee y \notin M$. So by $\bigcap \gamma(A\perp x) \subseteq M$, $x \vee y \notin \bigcap \gamma(A\perp x)$. But since $x \in A$ and $y \in \bigcap \gamma_W([\![\neg x]\!])$, we get $x \vee y \in A \cap \bigcap \gamma_W([\![\neg x]\!]) = \bigcap \gamma(A\perp x)$, and we have again a contradiction. Q.E.D.

As the domain $A\perp$ of selection functions for A is finitely additive, subtractive and compact, all domain requirements mentioned in Lemma 4(a)–(d) are satisfied. We can further adapt the conditions (I′)–(IV′) to our needs. We now know that for all $x, y \in A - Cn(\emptyset)$, $A\perp x \cup A\perp y = A\perp (x \wedge y)$, and $\gamma(A\perp x) \cap A\perp y = \emptyset$ iff $y \in \bigcap \gamma(A\perp x)$. With $\mathscr{X} = A\perp$, the general conditions (I′)–(IV′) can therefore be transformed into the following conditions on selection functions for theories:

(I″) For all $x, y \in A - Cn(\emptyset)$, $\gamma(A\perp (x \wedge y)) \subseteq \gamma(A\perp x) \cup \gamma(A\perp y)$
(II″) For all $x, y \in A - Cn(\emptyset)$, $\gamma(A\perp x) \cap \gamma(A\perp y) \subseteq \gamma(A\perp (x \wedge y))$
(III″) For all $x, y \in A - Cn(\emptyset)$, if $y \in \bigcap \gamma(A\perp (x \wedge y))$, then $\gamma(A\perp x) \subseteq \gamma(A\perp (x \wedge y))$
(IV″) For all $x, y \in A - Cn(\emptyset)$, if $x \notin \bigcap \gamma(A\perp (x \wedge y))$, then $\gamma(A\perp x) \subseteq \gamma(A\perp (x \wedge y))$

From these conditions, it will be a rather short step to some interesting postulates for rational ("economical") theory contraction.

The Finite Case

A set B of propositions will be called *logically finite* (or *finite modulo Cn*, or simply *finite*) if Cn partitions B into finitely many cells. The finite case is much easier to handle than the general one. This is due to the fact that every selection function over a logically finite theory is ω-covering and even full.

Notice that for a theory A to be finite modulo a common logic Cn, it will be necessary that the underlying language has only finitely many atoms. For suppose

there are infinitely many atoms $p_1, p_2, p_3, \ldots$ in our language. Then for every proposition x from A there are infinitely many atoms p_i not occurring in x. Thus the infinitely many $(x \vee p_i)$'s which are all contained in A will be mutually nonequivalent, so A is not finite. Conversely, if only a finite number of nonequivalent propositional operators of bounded arity is definable in Cn, as in classical propositional logic or in modal logics with finitely many modalities, then the finiteness of the number of propositional atoms is also sufficient for the logical finiteness of theories A phrased in the language in question.

Given a logically finite theory A, it is clear that each of the following sets is finite: $A\perp x$, for every x, $A\perp$, and U_A.

A *representative* of a logically finite set of propositions B is a conjunction of representatives of all the equivalence classes under Cn of propositions in B.

Henceforth, we shall use the following notational convention. For sets of propositions $A, B, \ldots, M, N, \ldots, W, \ldots$ which are finite modulo Cn, the lower case letters $a, b, \ldots, m, n, \ldots, w, \ldots$ denote their representatives. For any two sets A and B such that $B \subseteq A$, $\overline{b}_A$ denotes the disjunction of representatives of those equivalence classes under Cn, the elements of which are in $Cn(A) - Cn(B)$. When we are dealing with a fixed theory A, we simply write $\overline{b}$ instead of $\overline{b}_A$. If $B = A$, then $\overline{b}_A$ is defined to be the falsity, $\perp$. We may call $\overline{b}_A$ the *corepresentative* of B (relative to A).

It is easy to verify in the finite case that for $M \in U_A$ and $W = \mathscr{W}(M)$, w is equivalent with $\neg a \wedge m$, and that for $W \in V_A$ and $M = \mathscr{M}(W)$, m is equivalent with $a \vee w$. This helps us to get the following useful

Lemma 6. *Let A be a finite theory and $M \in U_A$.*

(i) *If w is the representative of $\mathscr{W}(M)$ and $\overline{w}_L$ is the corepresentative of $\mathscr{W}(M)$ relative to the set of all propositions in L, then the following four propositions are equivalent: $\overline{m}_A$, $m \rightarrow a$, $\overline{w}_L$ and $\neg w$.*

(ii) $A\perp \overline{m} = \{M\}$.

(iii) *For all $M_1, \ldots, M_n \in U_A$, if $m \vdash m_1 \vee \cdots \vee m_n$, then $M = M_i$ for some i.*

Lemma 6(ii), together with Corollary 1(ii), shows that in the finite case every subset $\mathbf{M}$ of U_A equals $A\perp x$ for some proposition x, viz., for $x = \bigwedge\{\overline{m} : M \in \mathbf{M}\}$. This in turn means that all selection functions for a finite theory A are ω-covering and in fact full. They are not only complete (Δ-elementary) but even elementary (cf. AGM 1985, Observation 4.6).

Postulates for Contraction Functions

We now turn to a set of rationality criteria which has gained some prominence in the literature on belief change. The *basic AGM postulates* are given by ($\dot{-}1$) – ($\dot{-}6$), and the two *supplementary* ones are ($\dot{-}7$) and ($\dot{-}8$). For their motivation, see Gärdenfors (1988).

($\dot{-}1$) $A\dot{-}x$ is a theory

($\dot{-}2$) $A\dot{-}x \subseteq A$

(÷3) If $x \notin A$ then $A \div x = A$
(÷4) If $x \in A \div x$ then $\vdash x$
(÷5) $A \subseteq Cn((A \div x) \cup \{x\})$
(÷6) If $Cn(x) = Cn(y)$ then $A \div x = A \div y$
(÷7) $A \div x \cap A \div y \subseteq A \div (x \wedge y)$
(÷8) If $x \notin A \div (x \wedge y)$ then $A \div (x \wedge y) \subseteq$A$\div x$.

These postulates, and all following postulates, are understood as quantified over all theories A and all propositions x and y. It follows from (÷1) and (÷5) that $A \div x = A$ for every $x \in Cn(\emptyset)$. We introduce two new conditions.

(÷8r) $A \div (x \wedge y) \subseteq Cn(A \div x \cup A \div y)$
(÷8c) If $y \in A \div (x \wedge y)$, then $A \div (x \wedge y) \subseteq A \div x$.

With two very recent exceptions,[5] I have never seen a condition like (÷8r) discussed in writings on the logic of theory change. Given (÷4), (÷8) implies the "covering condition"

$$A \div (x \wedge y) \subseteq A \div x \text{ or } A \div (x \wedge y) \subseteq A \div y$$

(AGM 1985, Observation 3.4) and hence (÷8r).[6] Postulate (÷8c) was found to be relevant in Rott (1992b), where it has the same name. The "r" in (÷8r) stands for "relational", and "c" stands for "cumulative". The first name will be explained by the present paper, the second one is explained in Rott (1992b). For contraction functions satisfying (÷4), (÷8c) is also a weakening of (÷8). However, there is no logical relationship between (÷8c) and (÷8r), not even in the finite case and in the presence of (÷1) – (÷7). To see this, consider the propositional language L over the two atoms p and q. In the following two figures, "$\binom{x}{y}$" is short for "$A \div x = Cn(y)$". It is easily verified that the contraction function ÷ over $A = Cn(p \wedge q)$ defined in Fig. 16.1 satisfies (÷1) – (÷7) and (÷8r), but it does not satisfy (÷8c), because $q \in A \div (p \wedge q)$ but $Cn(q) = A \div (p \wedge q) \not\subseteq A \div p = Cn(\neg p \vee q)$. On the other hand, the contraction function ÷ over the same theory A defined in Fig. 16.2 satisfies (÷1) – (÷7) and (÷8c), but it does not satisfy (÷8r), because $Cn(p \vee q) = A \div (p \wedge q) \not\subseteq Cn((A \div p) \cup (A \div q)) = Cn(\{\neg p \vee q, p \vee \neg q\}) = Cn(p \leftrightarrow q)$.

We now relate the abstract postulates for contraction functions to our previous requirements for selection functions in partial meet contraction.

[5]Both were discovered independently and concern belief revision rather than belief contraction. The first exception is condition (R8) in Katsuno and Mendelzon (1991). The second is the infinitary condition "Gamma" in Lindström (1991) which is labelled (BC7) in its variant for belief revision operations.

[6]Incidentally, it is proved in AGM (1985, Observation 6.5) that the conjunction of (÷7) and (÷8) is equivalent to the even stronger "ventilation condition"

$$A \div (x \wedge y) = A \div x \text{ or } A \div (x \wedge y) = A \div y \text{ or } A \div (x \wedge y) = A \div x \cap A \div y.$$

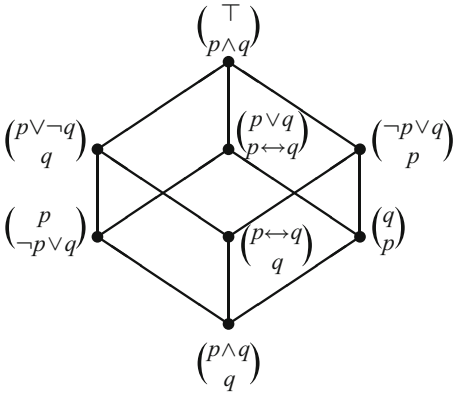


Fig. 16.1 Contraction function satisfying (÷1)–(÷7) and (÷8r), but not (÷8c)

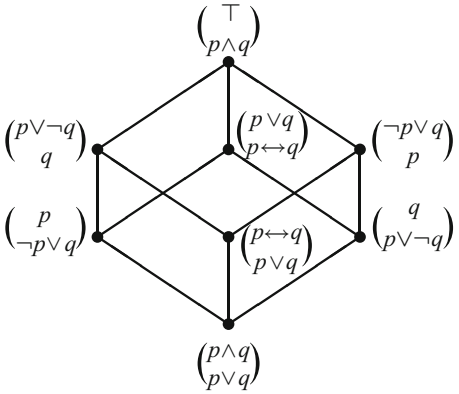


Fig. 16.2 Contraction function satisfying (÷1)–(÷7) and (÷8c), but not (÷8r)

Lemma 7. *Let A be a theory. If $\div = \mathscr{C}(\gamma)$ for some selection function γ for A and γ satisfies* $\left\{\begin{array}{ll}(a) & -\\ (b) & (I'')\\ (c) & (II'') \text{ and } \gamma \text{ is complete}\\ (d) & (III'')\\ (e) & (IV'')\end{array}\right\}$ *then $\div$ satisfies* $\left\{\begin{array}{ll}(a) & (\div 1)-(\div 6)\\ (b) & (\div 1)-(\div 7)\\ (c) & (\div 1)-(\div 6) \text{ and } (\div 8r)\\ (d) & (\div 1)-(\div 6) \text{ and } (\div 8c)\\ (e) & (\div 1)-(\div 6) \text{ and } (\div 8)\end{array}\right\}$.

Proof. Let $\div$ be a partial meet contraction function over A determined by γ.

(a) It is proved in AGM (1985, Observation 2.5), that $\div$ satisfies (÷1) – (÷6).

It is easy to verify that ($\dot{-}7$) and ($\dot{-}8$), and thus also ($\dot{-}8r$) and ($\dot{-}8c$), are satisfied whenever one of the limiting cases $x \notin A - Cn(\emptyset)$ or $y \notin A - Cn(\emptyset)$ holds. In the rest of this proof, we always presume the principal case $x, y \in A - Cn(\emptyset)$.

(b) Now let γ satisfy (I″), i.e., let $\gamma(A\perp(x \wedge y)) \subseteq \gamma(A\perp x) \cup \gamma(A\perp y)$. Hence $\bigcap(\gamma(A\perp x)) \cap \bigcap(\gamma(A\perp y)) = \bigcap(\gamma(A\perp x) \cup \gamma(A\perp y)) \subseteq \bigcap \gamma(A\perp(x \wedge y))$, i.e., $A\dot{-}x \cap A\dot{-}y \subseteq A\dot{-}(x \wedge y)$. That is, $\dot{-}$ satisfies ($\dot{-}7$).

(c) Now let γ be complete and satisfy (II″). Let $z \in A\dot{-}(x \wedge y) = \bigcap \gamma(A\perp(x \wedge y))$. So by (II″), $z \in \bigcap(\gamma(A\perp x) \cap \gamma(A\perp y))$. Now suppose for reductio that $z \notin Cn(A\dot{-}x \cup A\dot{-}y)$. Then there is an $M \in A\perp z$ such that $(\bigcap \gamma(A\perp x)) \cup (\bigcap \gamma(A\perp y)) \subseteq M$. Since γ is complete, we get $M \in \gamma(A\perp x)$ and $M \in \gamma(A\perp y)$, so $M \in \gamma(A\perp x) \cap \gamma(A\perp y)$. But $z \notin M$, so $z \notin \bigcap(\gamma(A\perp x) \cap \gamma(A\perp y))$ which gives us a contradiction. So $\dot{-}$ satisfies ($\dot{-}8r$).

(d) Now let γ satisfy (III″). Since the antecedents of (III″) and ($\dot{-}8c$) are identical and $\gamma(A\perp x) \subseteq \gamma(A\perp(x \wedge y))$ entails $\bigcap \gamma(A\perp(x \wedge y)) \subseteq \bigcap \gamma(A\perp x)$, it is obvious that $\dot{-}$ satisfies ($\dot{-}8c$).

(e) Now let γ satisfy (IV″). By exactly the same argument as in (d), $\dot{-}$ satisfies ($\dot{-}8$). Q.E.D.

Given a contraction function $\dot{-}$ over A, we can derive from it a selection function $\gamma_{\dot{-}}$ for A. The idea is that an element of $A\perp x$ is a best element of $A\perp x$ if it includes everything which is included in the contraction of A with respect to x. So we define $\gamma_{\dot{-}}(A\perp x)$ as $\{M \in A\perp x : A\dot{-}x \subseteq M\}$, for every $x \in A - Cn(\emptyset)$. Instead of $\gamma_{\dot{-}}$, we shall also write $\mathscr{S}(\dot{-})$.

Lemma 8. *Let A be a theory. If $\dot{-}$ is a contraction function over A satisfying*

$$\left\{\begin{array}{ll} (a) & (\dot{-}1) - (\dot{-}6) \\ (b) & (\dot{-}1) - (\dot{-}7) \\ (c) & (\dot{-}1) - (\dot{-}6) \textit{ and } (\dot{-}8r) \\ (d) & (\dot{-}1) - (\dot{-}6) \textit{ and } (\dot{-}8c) \\ (e) & (\dot{-}1) - (\dot{-}6) \textit{ and } (\dot{-}8) \end{array}\right\}$$

then $\dot{-}$ is a partial meet contraction function determined by the selection function $\mathscr{S}(\dot{-})$, i.e., $\dot{-} = \mathscr{C}(\mathscr{S}(\dot{-}))$, and $\mathscr{S}(\dot{-})$ is complete and satisfies

$$\left\{\begin{array}{ll} (a) & - \\ (b) & (I'') \\ (c) & (II'') \\ (d) & (III'') \\ (e) & (IV'') \end{array}\right\}.$$

Proof. Let $\dot{-}$ satisfy ($\dot{-}1$) – ($\dot{-}6$).

(a) It is proved in (AGM 1985, Observation 2.5) that $\mathscr{S}(\dot{-})$ is a selection function for A and that $\dot{-}$ is determined by $\mathscr{S}(\dot{-})$. We add the check for well-definedness.

In the principal case $x, y \in A - Cn(\emptyset)$, $A\perp x = A\perp y$ entails that $Cn(x) = Cn(y)$; for if it were the case that $x \nvdash y$, say, then there would be an element in $A\perp y$ containing x, contrary to $A\perp x = A\perp y$. But from $Cn(x) = Cn(y)$ and $(\dot{-}6)$ it follows by the definition of $\mathscr{S}(\dot{-})$ that $\gamma(A\perp x) = \gamma(A\perp y)$, as desired. By construction, $\mathscr{S}(\dot{-})$ is complete for every contraction function $\dot{-}$. It remains to prove the various additional properties of $\mathscr{S}(\dot{-})$.

(b) Let $\dot{-}$ in addition satisfy $(\dot{-}7)$, and let $M \in \gamma_{\dot{-}}(A\perp (x\wedge y))$. We want to show that $M \in \gamma_{\dot{-}}(A\perp x)\cup\gamma_{\dot{-}}(A\perp y)$. By $M \in \gamma_{\dot{-}}(A\perp (x\wedge y))$, $A\dot{-}(x\wedge y) \subseteq M$, so by $(\dot{-}7)$, $A\dot{-}x \cap A\dot{-}y \subseteq M$. Now suppose for reductio that $M \notin \gamma_{\dot{-}}(A\perp x) \cup \gamma_{\dot{-}}(A\perp y)$, i.e., neither $A\dot{-}x \subseteq M$ nor $A\dot{-}y \subseteq M$. Take $z \in (A\dot{-}x) - M$ and $z' \in (A\dot{-}y) - M$. By $z, z' \in A$, $z, z' \notin M$ and the primeness of all elements of U_A, $z \vee z' \notin M$. But on the other hand, since $z \vee z' \in A\dot{-}x \cap A\dot{-}y \subseteq M$, we get $z \vee z' \in M$, so we have a contradiction. So $\mathscr{S}(\dot{-})$ satisfies (I″).

(c) Now let $\dot{-}$ satisfy $(\dot{-}1)$ – $(\dot{-}6)$ and $(\dot{-}8r)$, and let $M \in \gamma_{\dot{-}}(A\perp x) \cap \gamma_{\dot{-}}(A\perp y)$. We want to show that $M \in \gamma_{\dot{-}}(A\perp (x \wedge y))$. But as our hypothesis means that $A\dot{-}x\cup A\dot{-}y \subseteq M$ and M is a theory, we have $Cn(A\dot{-}x\cup A\dot{-}y) \subseteq M$, so by $(\dot{-}8r)$ $A\dot{-}(x \wedge y) \subseteq M$, so $M \in \gamma_{\dot{-}}(A\perp (x \wedge y))$. So $\mathscr{S}(\dot{-})$ satisfies (II″).

(d) Now let $\dot{-}$ satisfy $(\dot{-}1)$ – $(\dot{-}6)$ and $(\dot{-}8c)$, and let $y \in \bigcap \gamma_{\dot{-}}(A\perp (x \wedge y)) = A\dot{-}(x \wedge y)$ and $M \in \gamma_{\dot{-}}(A\perp x)$. The latter condition means that $A\dot{-}x \subseteq M$, so by $(\dot{-}8c)$ $A\dot{-}(x \wedge y) \subseteq M$, so $M \in \gamma_{\dot{-}}(A\perp (x \wedge y))$, as desired. So $\mathscr{S}(\dot{-})$ satisfies (III″).

(e) Now let $\dot{-}$ satisfy $(\dot{-}1)$ – $(\dot{-}6)$ and $(\dot{-}8)$. By exactly the same argument as the last one, $\mathscr{S}(\dot{-})$ satisfies (IV″). Q.E.D.

Representation Theorems for Contraction Functions

We get four representation theorems for partial meet contraction as a result of conjoining previously established lemmas. Our Theorems 1 and 4 are proven in AGM (1985) although our construction for obtaining the latter appears to be quite different from the one used there. Theorems 2 and 3 are new; the "completeness" half of the former is generally valid, the remainder of the new results hold only under various additional conditions. All these conditions are satisfied if A is finite. It should be noted that our proofs of the completeness halves use one and the same method in that they all work with Samuelson revealed preferences.

Theorem 1. *A contraction function $\dot{-}$ over A satisfies $(\dot{-}1)$ – $(\dot{-}6)$ if and only if it is a partial meet contraction function.*

Proof. See Lemmas 7(a) and 8(a) above, according to which in the cases indicated $\mathscr{C}(\gamma)$ satisfies $(\dot{-}1)$ – $(\dot{-}6)$ and $\mathscr{S}(\dot{-})$ is a selection function for A such that $\mathscr{C}(\mathscr{S}(\dot{-})) = \dot{-}$. (Theorem 1 is Observation 2.5 of AGM 1985.) Q.E.D.

It is clear from Lemma 8 that in the proof of the completeness half of Theorem 1 the determining selection function is chosen complete, but we do not think it is

necessary to state this in the theorem. The same comment applies to the following three representation theorems.

Theorem 2. *Every relational partial meet contraction function $\dot{-}$ over A satisfies ($\dot{-}$1) – ($\dot{-}$7), and if $\dot{-}$ is determined by a selection function that is both relational and complete (equivalently, Δ-elementary), then it satisfies ($\dot{-}$8r). Conversely, every contraction function $\dot{-}$ over A satisfying ($\dot{-}$1) – ($\dot{-}$7) and ($\dot{-}$8r) is a relational partial meet contraction function.*

Proof. For the first part, let $\dot{-}$ be a partial meet contraction function determined by a preference relation $\leq$. By Lemma 1(a), $\mathscr{S}(\leq)$ satisfies (I) and (II). Since $A\perp$ is subtractive, it also satisfies (I′) and (II′), by Lemma 2(a), and also (I″) and (II″), by reformulation. So by Lemma 7 (b), $\mathscr{C}(\mathscr{S}(\leq))$ satisfies ($\dot{-}$1) – ($\dot{-}$7), and by Lemma 7(c), it satisfies ($\dot{-}$8r), if $\mathscr{S}(\leq)$ is complete, i.e., Δ-elementary.

For the converse, let $\dot{-}$ satisfy ($\dot{-}$1) – ($\dot{-}$7) and ($\dot{-}$8r). By Lemma 8(b) and (c), $\mathscr{C}(\mathscr{S}(\dot{-})) = \dot{-}$ and $\mathscr{S}(\dot{-})$ satisfies (I″) and (II″), and also (I′) and (II′), by reformulation. Since $A\perp$ is subtractive, $\mathscr{S}(\dot{-})$ satisfies (I), by Lemma 4(a). Since $A\perp$ is also finitely additive and compact, $\mathscr{S}(\dot{-})$ is relational with respect to $\mathscr{P}(\mathscr{S}(\dot{-}))$, by Lemma 4(b). That is, $\mathscr{S}(\dot{-}) = \mathscr{S}(\mathscr{P}(\mathscr{S}(\dot{-})))$. Hence $\dot{-} = \mathscr{C}(\mathscr{S}(\mathscr{P}(\mathscr{S}(\dot{-}))))$, i.e., $\dot{-}$ is relational with respect to $\mathscr{P}(\mathscr{S}(\dot{-}))$. Q.E.D.

The completeness of γ is necessary in the "soundness" part of Theorem 2. The possible failure without completeness is due to the fact that for sets W' and W'' of maximally consistent sets of propositions, it does not necessarily hold that $\bigcap (W' \cap W'') \subseteq Cn\,(\bigcap W' \cup \bigcap W'')$.

Theorem 3. *Every relational partial meet contraction function $\dot{-}$ over A determined by a smooth and negatively transitive preference relation satisfies ($\dot{-}$1) – ($\dot{-}$7) and ($\dot{-}$8c). In partial converse, every contraction function $\dot{-}$ over A satisfying ($\dot{-}$1) – ($\dot{-}$7), ($\dot{-}$8r), and ($\dot{-}$8c) is a relational partial meet contraction function, and if $A\perp$ is 123-covering, then the determining preference relation can be chosen negatively transitive.*

Proof. For the first part, let $\dot{-}$ be a partial meet contraction function determined by a smooth and negatively transitive preference relation $\leq$. We show in the same way as in the proof of Theorem 2 that $\mathscr{S}(\leq)$ satisfies (I′) and (II′), and we know from Lemmas 3(b) and 4(c) that it satisfies (III) and (III′). So by Lemma 7(b) and (d), $\mathscr{C}(\mathscr{S}(\leq))$ satisfies ($\dot{-}$1) – ($\dot{-}$7) and ($\dot{-}$8c).

For the converse, let $\dot{-}$ satisfy ($\dot{-}$1) – ($\dot{-}$7), ($\dot{-}$8r), and ($\dot{-}$8c). As in the proof of Theorem 2, we get that $\dot{-}$ is relational with respect to $\mathscr{P}(\mathscr{S}(\dot{-}))$. Moreover, by Lemma 8(b), $\mathscr{S}(\dot{-})$ satisfies (I′), and by Lemma 8(d), $\mathscr{S}(\dot{-})$ satisfies (III′) and also (III), by Lemma 4(c). Hence if $A\perp$ is 123-covering, then $\mathscr{P}(\mathscr{S}(\dot{-}))$ is negatively transitive, by Lemma 2(b). Q.E.D.

Theorem 4. *A contraction function $\dot{-}$ over A satisfies ($\dot{-}$1) – ($\dot{-}$6), ($\dot{-}$7), and ($\dot{-}$8) if and only if it is a transitively relational partial meet contraction function.*

Proof. This is Observation 4.4 of AGM (1985), but we shall sketch our proof which is quite different from the construction offered there.

For the first part of the theorem, let $\dot{-}$ be a partial meet contraction function determined by a transitive preference relation $\leq$. We show in the same way as in the proof of Theorem 2 that $\mathscr{S}(\leq)$ satisfies (I′) and (II′), and we know from Lemmas 3(c) and 4(d) that it satisfies (IV) and (IV′). So by Lemma 7(b), $\mathscr{C}(\mathscr{S}(\leq))$ satisfies $(\dot{-}1)$ – $(\dot{-}7)$, and by Lemma 7(e), it satisfies $(\dot{-}8)$.

For the converse, let $\dot{-}$ satisfy $(\dot{-}1)$ – $(\dot{-}7)$ and $(\dot{-}8)$. As in the proof of Theorem 2, we get that $\dot{-}$ is relational with respect to $\mathscr{P}(\mathscr{S}(\dot{-}))$. Moreover, by Lemma 8(e), $\mathscr{S}(\dot{-})$ satisfies (IV′) and also (IV), by Lemma 4(d). Since $A\perp$ is finitely additive, $\mathscr{P}(\mathscr{S}(\dot{-}))$ is transitive, by Lemma 2(c). Q.E.D.

Corollary 2. *Let A be a logically finite theory. Then* $\dot{-}$ *is a* $\left\{\begin{array}{c}\text{—}\\ \textit{relational}\\ \textit{negatively transitively relational}\\ \textit{transitively relational}\end{array}\right\}$ *partial meet contraction function iff it satisfies* $\left\{\begin{array}{c}(\dot{-}1)-(\dot{-}6)\\ (\dot{-}1)-(\dot{-}7)\ \textit{and}\ (\dot{-}8r)\\ (\dot{-}1)-(\dot{-}7),\ (\dot{-}8r)\ \textit{and}\ (\dot{-}8c)\\ (\dot{-}1)-(\dot{-}7)\ \textit{and}\ (\dot{-}8)\end{array}\right\}$.

Proof. By Theorems 1, 2, 3, and 4, because for finite theories A, γ over $A\perp$ is complete, every negatively transitive nonstrict preference relation $\leq$ over A is smooth, and $A\perp$ is 123-covering. □

It is easy to locate the impact of the finiteness assumption. In the case of relational partial meet contraction s (Theorem 2), it is a constraint (completeness) imposed by model theory which has to be met in order to make the "soundness" direction work. In the case of negatively transitively relational partial meet contraction s (Theorem 3), it is a constraint imposed by the theory of rational choice (that γ be 123-covering and smooth) which is satisfied without strains on intuitive plausibility only in the finite case.

It should be noted that although $(\dot{-}8)$ in the context of the other postulates implies $(\dot{-}8r)$ and $(\dot{-}8c)$, transitive relationality does not imply negatively transitive relationality. However, transitivity in companion with connectivity implies negative transitivity. And it is known from AGM (1985, § 5) that a connectivity requirement on the underlying preference relation changes very little in the partial meet mechanism.[7]

We conclude this section with a direct representation of the Samuelson preferences over U_A revealed by contraction behavior. Let $\leq = \mathscr{P}(\mathscr{S}(\dot{-}))$. Then in the general case, $M \leq M'$ iff there is an x such that $A\dot{-}x \subseteq M'$ and $M \in A\perp x$, or there

[7] In order to reconcile this with your intuitions, cf. Footnote 1.

is no x such that $A \dot{-} x \subseteq M$. We recall, for contrast, the construction used in the proof of AGM (1985, Observation 4.4). There $M \leq M'$ is defined to hold between sets M and M' from U_A iff for all x such that $A \dot{-} x \subseteq M$ and $M' \in A \perp x$, it holds that $A \dot{-} x \subseteq M'$, and there is an $x \in A$ such that $A \dot{-} x \subseteq M'$.[8] Evidently, this definition is quite different from the one we used, and it is doubtful whether it is suitable for any case but the strongest one which is treated in Theorem 4. It is interesting, however, that AGM employ our construction in their proof of Observation 4.10 of AGM (1985). In the finite case, it is not difficult to find out that according to our definition, $M \leq M'$ iff $A \dot{-} \overline{m} \wedge \overline{m'} \subseteq M'$, or equivalently(!), iff $\overline{m'} \notin A \dot{-} \overline{m} \wedge \overline{m'}$.

The Logic of Prioritized Base Contractions in the Finite Case

In this section we present the concept of prioritized base contraction as studied in particular by Nebel (1989, 1992). On the face of it, these belief change operations are entirely different from the partial meet contractions discussed so far. As Nebel has shown, however, very close connections can be established between the two approaches. We shall further elaborate this theme and determine the logic of prioritized base contractions in terms of the above-mentioned postulates by applying one of our central representation theorems.

Let B be a finite set of propositions. B is called a *(simple) belief base* for the theory A if $Cn(B) = A$. B is called *prioritized* if there is a weak ordering, i.e., a transitive and connected relation $\preceq$ over B (Nebel's 1989, 1992 "epistemic relevance"). The relation $\sim = \preceq \cap \preceq^{-1}$ partitions B into finitely many equivalence classes $B_1, \ldots, B_n$. We label the classes in such a way that $x_i \preceq x_j$ for $x_i \in B_i$ and $x_j \in B_j$ with $i \leq j$. Generalizing slightly, we say that a *prioritized belief base* for A is a finite sequence $\langle B_1, \ldots, B_n \rangle$, or $\langle B_i \rangle$ for short, of finite sets of propositions such that $Cn(\bigcup B_i) = A$. Simple belief bases are now special cases of prioritized belief bases with $n = 1$. We do not require that the B_i's are disjoint or mutually consistent. Intuitively, the propositions in B_n are the most important (most relevant, most valuable, most certain, etc.) pieces of information, and the propositions in B_1 are the least important. One can think of the complement of $\bigcup B_i$ as the set B_0 of propositions for which there is no independent evidence. Although the finiteness of $B = \bigcup B_i$ does not ensure that the generated theory A is logically finite, our central result in this section, Theorem 7, will be restricted to this case.

Prioritized base contractions are contractions of theories A presented in the form of (generated by, associated with) a fixed prioritized belief base $\langle B_i \rangle$. The basic idea is to minimize the loss of propositions at every level of "priority".

For every $i \in \{1, \ldots, n\}$, let $B_{\geq i}$ be the set of all elements of $B = \bigcup B_i$ with a priority of at least i, i.e., $B_{\geq i} = \bigcup \{B_j : j \geq i\}$. Furthermore, let $C_i = C \cap B_i$ and

[8] This simplified rephrasing of the AGM construction makes use of the fact that for all $M \in U_A$ and contraction functions $\dot{-}$ satisfying the AGM postulates, $A \dot{-} x \subseteq M$ implies $M \in A \perp x$.

$C_{\geq i} = C \cap B_{\geq i}$ for any set of propositions C. Then, in the context of this section, C is a *preferred x-discarding* subset of B if $C \subseteq B$ and for every i, $C_{\geq i}$ is an inclusion maximal subset of $B_{\geq i}$ subject to the condition that x is not implied.

Two kinds of information are used in prioritized base contraction: the syntactical information encoded in the structure of the propositions in B, and the weighting of these propositions expressed by $\preceq$. If the belief base is simple, i.e., if $n = 1$, then we only exploit syntactical information. Prioritized base contractions and revisions are studied in an infinitistic framework by Nebel (1989, 1992).[9] He shows that prioritized base revisions can be represented by partial meet theory revisions. Then he proves that a certain variant of simple base contraction satisfies ($\dot{-}1$) – ($\dot{-}7$) but not ($\dot{-}8$), and a corresponding result for prioritized base revisions. However, Nebel leaves open the question which logic is *characterized* by prioritized base contractions. Building on Theorem 3 and its corollary, we shall answer this question for the finite case. Our concern will be the slightly more complicated case of prioritized base contractions, and we try to make our proof more transparent by introducing the concept of a base-oriented selection function.

Now consider the following strict preference relations between arbitrary sets of propositions. For every $i \in \{1, \ldots, n\}$ and any two sets C and C', we write $C \lessdot_i C'$ if and only if $C_i \subset C'_i$ and $C_j \subseteq C'_j$ for every $j > i$. We write $C \lessdot C''$ if there is an i such that $C \lessdot_i C'$. The relation $\lessdot$ is to reflect the fact that intuitively a set C' is better than another set C just in case it contains more important beliefs than C. In particular, $\lessdot$ satisfies a version of Hansson's (1992) maximizing property, because $C \cap B \subset C' \cap B$ implies $C \lessdot C'$. It is immediately verified that C is a preferred x-discarding subset of B if and only if it is a maximal element in $B \perp x$ under $\lessdot$. (Recall that $B \perp x$ is the set $\{N \subseteq B : x \notin Cn(N)$ and $x \in Cn(D)$ for all D with $N \subset D \subseteq B\}$.)

Following Nebel, we let $\underline{\lessdot}$ be the converse complement of $\underline{\lessdot}$, i.e., $C \underline{\lessdot} C'$ iff not $C' \lessdot C$. We denote this preference relation over arbitrary sets of propositions by $\mathscr{P}(\langle B_i \rangle)$. Clearly, since $\lessdot$ is irreflexive, asymmetric, conversely well-founded and transitive, $\mathscr{P}(\langle B_i \rangle)$ is reflexive, connected, negatively well-founded and negatively transitive. But $\mathscr{P}(\langle B_i \rangle)$ is not transitive, not even in the special case of simple belief bases ($n = 1$) where $\underline{\lessdot}$ coincides with $\not\supset$.

As before, a selection function $\gamma = \mathscr{S}(\underline{\lessdot})$ can be defined by stringent maximization. But this time γ is a selection function for the base B as well as for the theory A, and its domain may in fact be construed as the class of all nonempty sets of sets propositions in L, that is $2^{2^L} - \{\emptyset\}$. In the following, it is understood that $\gamma(B \perp x) = \{N \in B \perp x : N' \underline{\preceq} N$ for all $N' \in B \perp x\}$, while $\gamma(A \perp x) = \{M \in A \perp x : M' \underline{\lessdot} M$ for all $M' \in A \perp x\}$. Either way, γ is a selection function.

[9] Nebel's (1992) treatment of the fully general infinite case is not quite correct. Slips have crept into his claim that his $C \Downarrow \phi$ is nonempty, into his definition (9) of $\ll$, and into the proof of Proposition 8. As Nebel (personal communication) has suggested, they can be remedied by imposing a condition of converse well-foundedness on $\preceq$.

Let $\langle B_i \rangle$ be a prioritized belief base for A, and let $\gamma = \mathscr{S}(\mathscr{P}(\langle B_i \rangle))$ be the selection function over $A\perp$ determined by $\langle B_i \rangle$. A straightforward idea to get prioritized base contractions is the following:

$$A \dot{-} x = \begin{cases} \bigcap \{Cn(N) : N \in \gamma(B\perp x)\} & \text{for every } x \in A - Cn(\emptyset) \\ A & \text{for every } x \notin A \text{ and every } x \in Cn(\emptyset) \end{cases}$$

In the special case of simple belief bases ($n = 1$) we have $\gamma(B\perp x) = B\perp x$, so this definition boils down to the *full* meet $\bigcap\{B\perp x\}$.

Contraction functions obtained in this way satisfy most but not all of the postulates we would want them to satisfy.

Theorem 5. *If $\dot{-}$ is a prioritized base contraction function as determined by the above definition, then it satisfies ($\dot{-}$1) – ($\dot{-}$4), ($\dot{-}$6), ($\dot{-}$7), and ($\dot{-}$8c). However, even if the base is simple, $\dot{-}$ will in general fail to satisfy ($\dot{-}$5) and ($\dot{-}$8r).*

Proof. It is obvious that $\dot{-}$ satisfies ($\dot{-}$1) – ($\dot{-}$4) and ($\dot{-}$6).

For ($\dot{-}$7), assume that $N \vdash z$ for every $N \in \gamma(B\perp x) \cup \gamma(B\perp y)$. We need to show that $N' \vdash z$ for every $N' \in \gamma(B\perp (x \wedge y))$. Let $N' \in \gamma(B\perp (x \wedge y)) \subseteq B\perp (x \wedge y)$. First we note that every element of $B\perp (x \wedge y)$ is either in $B\perp x$ or in $B\perp y$. Without loss of generality, assume that $N' \in B\perp x$. We are ready if we can show that N' is in $\gamma(B\perp x)$. Suppose it is not. Then there is an $N_1 \in B\perp x$ such that $N' \lll N_1$. N_1 must not be in $B\perp (x \wedge y)$, since N' is in $\gamma(B\perp (x \wedge y))$. Because $N_1 \in B\perp x - B\perp (x \wedge y)$, we get that there is a proper superset N_2 of N_1 such that $N_2 \in B\perp y$. Since $N_1 \in B\perp x$, $N_2 \vdash x$. Since every proper superset of N_2 satisfies both x and y, N_2 is in $B\perp (x \wedge y)$. By the maximizing property, we have $N_1 \lll N_2$. By the transitivity of $\lll$, we get $N' \lll N_2$. This, however, contradicts $N' \in \gamma(B\perp (x \wedge y))$ and $N_2 \in B\perp (x \wedge y)$. Therefore our supposition must have been false.

For ($\dot{-}$8c), assume that, first, $N \vdash y$ and, second, $N \vdash z$ for every $N \in \gamma(B\perp (x \wedge y))$. We need to show that $N' \vdash z$ for every $N' \in \gamma(B\perp x)$. Let $N' \in \gamma(B\perp x)$ and suppose for reductio that $N' \nvdash z$. If N' is in $B\perp (x \wedge y)$ then there must be, by the second assumption, an N_1 in $B\perp (x \wedge y)$ such that $N' \lll N_1$. In fact, N_1 can be chosen from $\gamma(B\perp (x \wedge y))$, by the transitivity and converse well-foundedness of $\lll$. By the first assumption, $N_1 \in B\perp x$. But this contradicts $N' \lll N_1$ and $N' \in \gamma(B\perp x)$. Hence N' cannot be in $B\perp (x \wedge y)$. But because $N' \nvdash x \wedge y$, there must be a proper superset N_2 of N' in $B\perp (x \wedge y)$. We know that $N_2 \vdash x$, by $N' \in B\perp x$. By the first assumption then, N_2 cannot be in $\gamma(B\perp (x \wedge y))$. By the transitivity and converse well-foundedness of $\lll$, there must be an N_3 in $\gamma(B\perp (x \wedge y))$ such that $N_2 \lll N_3$. By our first assumption, N_3 is in $B\perp x$. By the maximizing property, we have $N' \lll N_2 \lll N_3$, so by transitivity $N' \lll N_3$. This, however, contradicts $N' \in \gamma(B\perp x)$ and $N_3 \in B\perp x$. Therefore our supposition must have been false.

The failure of ($\dot{-}$5) is obvious. Consider for instance $B = \{p \wedge q\}$, for which $A = Cn(p \wedge q) \not\subseteq Cn((A \dot{-} p) \cup \{p\}) = Cn(p)$.

The failure of ($\dot{-}$8r) is not so easy to see. The simplest examples involve three atoms p, q, r and look somewhat like the following. Let B consist of the following four propositions:

$$(p \wedge q) \vee (p \wedge r) \vee (q \wedge r), \quad r \rightarrow (p \leftrightarrow q),$$
$$q \wedge (p \leftrightarrow r), \quad p \wedge r.$$

Clearly $A = Cn(B) = Cn(p \wedge q \wedge r)$. Note that $q \wedge (p \leftrightarrow r)$ implies $r \rightarrow (p \leftrightarrow q)$ and that $p \wedge r$ implies $(p \wedge q) \vee (p \wedge r) \vee (q \wedge r)$. $(p \wedge q) \vee (p \wedge r) \vee (q \wedge r)$ and $r \rightarrow (p \leftrightarrow q)$ taken together imply $p \wedge q$. This gives us

$$B\perp p = \{\{(p \wedge q) \vee (p \wedge r) \vee (q \wedge r)\}, \{r \rightarrow (p \leftrightarrow q), q \wedge (p \leftrightarrow r)\}\}$$

and

$$B\perp q = \{\{r \rightarrow (p \leftrightarrow q)\}, \{(p \wedge q) \vee (p \wedge r) \vee (q \wedge r), p \wedge r\}\}.$$

Therefore, the above account yields $A\dot{-}p = Cn(q \vee (p \wedge r))$ and $A\dot{-}q = Cn(\neg(\neg p \wedge q \wedge r))$, so $Cn(A\dot{-}p \cup A\dot{-}q) = Cn((p \wedge r) \vee (q \wedge \neg r))$. On the other hand,

$$B\perp (p \wedge q) = \{\{(p \wedge q) \vee (p \wedge r) \vee (q \wedge r), p \wedge r\}, \{r \rightarrow (p \leftrightarrow q), q \wedge (p \leftrightarrow r)\}\},$$

yielding $A\dot{-}p \wedge q = Cn((p \wedge r) \vee (\neg p \wedge q \wedge \neg r))$. At last we realize that the union of $A\dot{-}p$ and $A\dot{-}q$ does not imply $(p \wedge r) \vee (\neg p \wedge q \wedge \neg r)$, because the former, but not the latter is compatible with $p \wedge q \wedge \neg r$. Q.E.D.

In so far as one considers the postulates ($\dot{-}5$) and ($\dot{-}8r$) to be plausible principles of theory change—which in fact we do—, Theorem 5 shows that the above definition is defective. For this reason we introduce a slightly more sophisticated concept, and our official definition reads as follows. Let a again be a representative of A.

The *prioritized base contraction* $\dot{-}$ over a logically finite theory A determined by $\langle B_i \rangle$ is then given by

$$A\dot{-}x = \begin{cases} \bigcap \{Cn(N \cup \{x \rightarrow a\}) : N \in \gamma(B\perp x)\}^{10} & \text{for every } x \in A - Cn(\emptyset) \\ A & \text{for every } x \notin A \text{ and } x \in Cn(\emptyset) \end{cases}$$

The singleton $\{x \rightarrow a\}$ is inserted in order to make sure that the recovery postulate ($\dot{-}5$) is satisfied. This trick is used by Nebel (1989).[11] We shall see that this modification also solves our problem with ($\dot{-}8r$). As a first indication

[10] Or equivalently, $Cn((\bigcap\{Cn(N) : N \in \gamma(B\perp x)\}) \cup \{x \rightarrow a\}) = Cn(A\dot{-}'x \cup \{x \rightarrow a\})$ where $A\dot{-}'x$ follows the first definition.

[11] Nebel's (1992) later paper deals with revisions where this problem does not arise. In fact, if revisions are construed as being generated by the so-called Levi-identity $A * x = Cn((A\dot{-}\neg x) \cup \{x\})$, then the modification made in our official definition does not have any effect on revisions.

of this, note that the official definition cures our above counterexample to ($\dot{-}8r$) by strengthening $A\dot{-}x$ to $Cn(y \wedge (x \rightarrow z))$ and $A\dot{-}y$ to $Cn(y \rightarrow (x \wedge z))$, so that $Cn(A\dot{-}x \cup A\dot{-}y) = Cn(x \wedge y \wedge z)$. Since this has been the full theory A, ($\dot{-}8r$) is clearly satisfied.

We denote the prioritized base contraction over A determined by the prioritized belief base $\langle B_i \rangle$ by $\mathscr{C}(\langle B_i \rangle)$. A contraction function $\dot{-}$ over A is a *simple* (or *prioritized*) *base contraction function* if there is a simple belief base B (a prioritized belief base $\langle B_i \rangle$) for A such that $\dot{-} = \mathscr{C}(B)$ (respectively, $\dot{-} = \mathscr{C}(\langle B_i \rangle)$). In conformity with the terminology of AGM (1985), simple base contractions could also be named *full meet* base contractions.

Lemma 9. *Let* $B = \bigcup B_i$, $A = Cn(B)$, $M \in A\perp x$, *and let* $\gamma = \mathscr{S}(\mathscr{P}(\langle B_i \rangle))$. *Then* $M \in \gamma(A\perp x)$ *iff* $N \subseteq M$ *for some* $N \in \gamma(B\perp x)$.

Let us call selection functions which harmonize choices in $A\perp$ with choices in $B\perp$ in the way exhibited in Lemma 9 base-oriented. The following result shows that the selection function γ induced by a prioritized belief base leads to the same result when applied directly to belief bases as when applied to the generated theory. Prioritized base contraction thus reduces to a special kind of partial meet theory contraction. A proof of almost the same fact has already been given by Nebel (1989, Theorem 14, 1992, Theorem 7). Our proof is somewhat more general in that it only turns on the fact that γ is base-oriented. Nothing hinges on the particular definition of $\underline{\ll}$.

Theorem 6. *Let* $\langle B_i \rangle$ *be a prioritized belief base for a logically finite theory A, and let* $\gamma = \mathscr{S}(\mathscr{P}(\langle B_i \rangle))$. *Then* $\mathscr{C}(\langle B_i \rangle) = \mathscr{C}(\gamma)$.

Proof. For the principal case $x \in A - Cn(\emptyset)$, we have to show that

$$\bigcap \{Cn(N \cup \{x \rightarrow a\}) : N \in \gamma(B\perp x)\} = \bigcap \gamma(A\perp x).$$

By Lemma 9, we know that γ is base-oriented.

"LHS $\subseteq$ RHS": Let $N \cup \{x \rightarrow a\} \vdash y$ for every $N \in \gamma(B\perp x)$, and let $M \in \gamma(A\perp x)$. Since γ is base-oriented, there is an $N \in \gamma(B\perp x)$ such that $N \subseteq M$. So $M \cup \{x \rightarrow a\} \vdash y$. But also, by the fullness of the elements in $A\perp x$, $x \rightarrow a \in M$. So $M \vdash y$, as desired.

"RHS $\subseteq$ LHS": Assume that there is an $N \in \gamma(B\perp x)$ such that $N \cup \{x \rightarrow a\} \nvdash y$. Hence, by the deduction theorem and supraclassicality, $N \cup \{y \rightarrow a\} \nvdash x$. Thus, there is an $M \in A\perp x$ such that $N \cup \{y \rightarrow a\} \subseteq M$. Since $N \subseteq M$ and γ is base-oriented, $M \in \gamma(A\perp x)$. Since $y \rightarrow a \in M$, $a \vdash x$ and $M \nvdash x$, we have $y \notin M$. So $y \notin \bigcap \gamma(A\perp x)$, as desired. Q.E.D.

We are now going to construct, for an arbitrary given preference relation $\leq$ over a logically finite theory A, an equivalent simple belief base B for A. This base will be denoted $\mathscr{B}(\leq)$. For $M \in U_A$, let $\hat{M}$ be the set $\{M' \in U_A : M' \not\leq M\} \cup \{M\}$ of elements in U_A which are not covered by M, together with M itself. Let $\hat{m}$ be an abbreviation for the disjunction $\bigvee\{m' : M' \in \hat{M}\}$, with each m' being a representative of the

respective M'. Then $B = \mathscr{B}(\leq)$ is defined as the set $\{\hat{m} : M \in U_A\}$. Since A is finite modulo Cn, everything else is. This construction and the proof idea for the direction "(iii)⇒(i)" in Theorem 7 is adapted from Lewis (1981).

Lemma 10. *Let A be a logically finite theory and $\leq$ be a reflexive and negatively transitive preference relation over U_A. Then $\mathscr{P}(\mathscr{B}(\leq)) = \leq$.*

Proof. As a preparatory consideration, let us consider what is contained in a set $M \cap B$, when $M \in U_A$ and $B = \mathscr{B}(\leq)$ for a given $\leq$. By definition, $M = Cn(m)$. Now any proposition from B is of the form $\widehat{m^*}$ for some $M^* \in U_A$, or more explicitly, of the form $\bigvee\{m^{**} : M^{**} \in U_A$, and $M^{**} \not\leq M^*$ or $M^{**} = M^*\}$. We know that a proposition $\widehat{m^*}$ from B is in M if and only if $m \vdash \widehat{m^*}$, i.e., $m \vdash \bigvee\{m^{**} : M^{**} \in U_A$, and $M^{**} \not\leq M^*$ or $M^{**} = M^*\}$. By Lemma 6(iii), this obtains just in case $M \not\leq M^*$ or $M = M^*$.

Now let $\leq^* = \mathscr{P}(\mathscr{B}(\leq))$. We want to show that $M \leq^* M'$ iff $M \leq M'$. By definition, $M \leq^* M'$ if and only if either $M \cap B \subseteq M' \cap B$ or $M' \cap B \not\subseteq M \cap B$. Applying our preparatory consideration, this means that either

(i) For every $M^* \in U_A$, if $M \not\leq M^*$ or $M = M^*$, then $M' \not\leq M^*$ or $M' = M^*$, or
(ii) There is an $M^* \in U_A$ such that $M' \not\leq M^*$ or $M' = M^*$, and $M \leq M^*$ and $M \neq M^*$.

We need to show that this is equivalent with $M \leq M'$. That the latter implies (i) or (ii) is clear: If $M = M'$, then (i), and if $M \neq M'$, then (ii) is immediate by putting $M^* = M'$.

For the converse, suppose that $M \not\leq M'$. We verify that this is incompatible with each of (i) and (ii). From (i), we get by putting $M^* = M$ that $M' \not\leq M$ or $M' = M$. In the former case we get from $M \not\leq M'$ and negative transitivity that $M \not\leq M$, contradicting the reflexivity of $\leq$. The latter case doesn't fare better; it contradicts $M \not\leq M'$ as well. From (ii), we get that $M' \not\leq M^*$ or $M' = M^*$ for some M^*. In the former case, $M \not\leq M'$ and negative transitivity would give us $M \not\leq M^*$, contradicting (ii). In the latter case, $M \not\leq M'$ would equally well yield $M \not\leq M^*$, again contradicting (ii). Q.E.D.

Theorem 7. *Let $\dot{-}$ be a contraction function over a logically finite theory A. Then the following conditions are equivalent:*

(i) *$\dot{-}$ is a simple base contraction function;*
(ii) *$\dot{-}$ is a prioritized base contraction function;*
(iii) *$\dot{-}$ satisfies ($\dot{-}1$) – ($\dot{-}7$), ($\dot{-}8r$), and ($\dot{-}8c$).*

Proof. That (i) entails (ii) is trivial.

To show that (ii) entails (iii), observe that by Theorem 6, $\mathscr{C}(\langle B_i\rangle) = \mathscr{C}(\mathscr{S}(\mathscr{P}(\langle B_i\rangle)))$ and that $\mathscr{P}(\langle B_i\rangle)$ is reflexive and negatively transitive. So by Corollary 2, $\mathscr{C}(\langle B_i\rangle)$ satisfies ($\dot{-}1$) – ($\dot{-}7$), ($\dot{-}8r$), and ($\dot{-}8c$).

To show that (iii) entails (i), we conclude from (iii) with the help of Corollary 2 that there is a negatively transitive preference relation $\leq$ such that $\dot{-} = \mathscr{C}(\mathscr{S}(\leq))$.

We know from Lemma 10 that the latter is identical with $\mathscr{C}(\mathscr{S}(\mathscr{P}(\mathscr{B}(\leq))))$. So there is a simple belief base B, namely $B = \mathscr{B}(\leq)$, such that $\dot{-} = \mathscr{C}(\mathscr{S}(\mathscr{P}(B)))$. By Theorem 6 then, $\dot{-} = \mathscr{C}(B)$. Q.E.D.

Theorem 7 tells us that for every prioritized belief base $\langle B_i \rangle$ for A there is a simple belief base B for A such that $\mathscr{C}(B) = \mathscr{C}(\langle B_i \rangle)$. In the finite case, every prioritization can be encoded syntactically. Does this mean that prioritization is superfluous? In answering this question we first have to emphasize that our generation of B from $\langle B_i \rangle$ took a rather roundabout route: $B = \mathscr{B}(\mathscr{P}(\langle B_i \rangle))$. An interesting problem now is whether a more perspicuous construction of B from $\langle B_i \rangle$ is possible. This question, too, is put in informal terms, and as such it permits no absolutely precise answer. Still we think that the answer must be no. Even for the most straightforward prioritized belief bases $\langle B_i \rangle$ the generated simple base $B = \mathscr{B}(\mathscr{P}(\langle B_i \rangle))$ becomes grossly unintuitive, and there is no prospect of finding different solutions to the problem. Consider for example the base containing p and $p \to q$ with the alternative prioritizations $\langle \{p \to q\}, \{p\} \rangle$ and $\langle \{p\}, \{p \to q\} \rangle$. In the former case, $B = \{p \wedge q, p, p \vee q, q \to p\}$, while in the latter case, $B' = \{q, p \leftrightarrow q\}$ will lead to exactly the same results as the prioritized belief base. But in neither case is there anything like a transparent relation to the original base $\langle B_i \rangle$. It appears that prioritization is useful, notwithstanding its formal dispensability in the finite case.

Acknowledgements I gratefully acknowledge numerous comments and suggestions by David Makinson which have once again been extremely helpful.

Appendix 1: Relating Theory Change and Nonmonotonic Logic

In a recent paper, Makinson and Gärdenfors (1991) make considerable progress toward linking the areas of theory change and nonmonotonic reasoning. They close with the following problem which we quote in full:

> We end with an important open question. In their (1985), Alchourrón, Gärdenfors and Makinson established a representation theorem for theory revision operations $*$ satisfying conditions $(*1)$–$(*8)$, in terms of "transitively relational partial meet revisions". The proof went via a representation theorem for a contraction function $\dot{-}$ satisfying certain conditions $(\dot{-}1)$ – $(\dot{-}8)$. On the other hand, Kraus et al. (1990) have established a representation theorem for supraclassical, cumulative and distributive nonmonotonic inference relations $\mid\sim$ defined between individual propositions, in terms of classical stoppered preferential model structures. The former proof is relatively short and abstract; the latter seems more complex. Also, the latter has not been generalized to a representation theorem for supraclassical, cumulative and distributive inference operations $C : 2^L \to 2^L$ [...] Does the representation theorem for theory change hold the key for a solution to this problem of extending the Kraus/Lehmann/Magidor representation theorem to the infinite case—despite the failure of consistency preservation for preferential model structures? Or do we have two essentially different representation problems? (Makinson and Gärdenfors 1991, pp. 203–204, notation adapted)

This question does not have a simple answer. Three different points have to be taken into consideration.

First, the approaches of AGM and KLM are not as distinct as Makinson and Gärdenfors seemed to assume. AGM contract and revise by single propositions, and similarly KLM consider the nonmonotonic consequences of simple propositions. Makinson and Gärdenfors's equation $y \in A * x$ iff $x \mathrel{|\!\sim} y$ (iff $y \in C(x)$) fully reflects this. A truly infinitistic stance towards both theory change and nonmonotonic logic is taken only by Lindström (1991). The question of whether the theory A is logically finite has no bearing on this issue. Here Makinson and Gärdenfors saw a difference which simply does not exist.

Secondly, Makinson and Gärdenfors tacitly passed over the fact that in KLM there is no counterpart to ($*8$) or ($\dot{-}8$). But this difference is indeed crucial, as is clear from a later paper of Lehmann and Magidor (1992). As regards the *preferential* logics dealt with by KLM, no "relatively short and abstract" proof seems to be possible. Thus it appears reasonable to construe Makinson and Gärdenfors's question as referring to the *rational* logics treated by Lehmann and Magidor.

But thirdly, Lehmann and Magidor's rational logics still differ from AGM-style theory revisions in that they are not required to satisfy a condition of consistency preservation which corresponds to postulate ($\dot{-}4$) for contractions. In this respect, Makinson and Gärdenfors show a keen sense of the intricacies of the situation. In unpublished notes, we have applied the techniques developed in this paper to the problem of providing rational logics with canonical models. We have found that it *is* possible to transfer our proof to the area of nonmonotonic logic, but that consistency preservation with respect to the underlying monotonic logic *Cn* is, in fact, indispensable for a perfect matching. The reason is that the compactness property presupposed for *Cn* runs idle if there are any "inaccessible worlds" (and this is just what a violation of consistency preservation amounts to). Since the results of our efforts bear considerable similarity with the venerable presentation in Lewis (1973)—except for the fact that the role of Lewis's extra-linguistic propositions (sets of "real" possible worlds) is played by the propositions of the object language—we had better refrain from expounding them here in more detail.

Appendix 2: Some Proofs and Examples[12]

Proof of Lemma 1. (a) Immediate from the definition of relationality. (If x is greatest in S' then it is so in all subsets of S' in which it is contained. If x is greatest in every S_i then it is so in $\bigcup S_i$.)

(b) Let γ be 12-covering and satisfy (I) and (II). We have to show that for every $S \in \mathscr{X}$, $\gamma(S) = \{x \in S :$ for all $y \in S$, $\{x, y\} \in \mathscr{X}$ and $x \in \gamma(\{x, y\})\}$. *LHS* $\subseteq$ *RHS*: Let $x \in \gamma(S)$ and $y \in S$. As γ is 12-covering, $\{x, y\} \in \mathscr{X}$ and (I) gives us $x \in \gamma(\{x, y\})$.

[12]This appendix was not included in the original 1993 publication of this paper.

RHS $\subseteq$ *LHS*: Let $x \in S$ and assume that for every $y \in S$, $\{x, y\} \in \mathscr{X}$ and $x \in \gamma(\{x, y\})$. Note that $\bigcup\{\{x, y\} : y \in S\} = S \in \mathscr{X}$ and $x \in \bigcup\{\{x, y\} : y \in S\}$. By (II), we get $x \in \gamma(\bigcup\{\{x, y\} : y \in S\}) = \gamma(S)$, as desired.

(c) Let γ be 12-covering and satisfy (I).

$\mathscr{P}_2(\gamma) \subseteq \mathscr{P}(\gamma)$: This direction is always valid.

$\mathscr{P}(\gamma) \subseteq \mathscr{P}_2(\gamma)$: Let $x \leq_\gamma y$. Then there is an $S \in \mathscr{X}$ such that $y \in \gamma(S) \subseteq S$ and $x \in S$. Because γ is 12-covering, $\{x, y\} \in \mathscr{X}$. But as $\{x, y\} \subseteq S$, (I) gives us $y \in \gamma(\{x, y\})$, so $x \leq_{\gamma,2} y$.

(d) Follows from (b) and (c).

(e) Follows from (a), (b), and (d).

(f) Let γ be additive and satisfy (I) and (II). We have to show that for every $S \in \mathscr{X}$, $\gamma(S) = \{x \in S :$ for all $y \in S$ there is a $T \in \mathscr{X}$ such that $x \in \gamma(T)$ and $y \in T\}$.

LHS $\subseteq$ *RHS*: Take $T = S$.

RHS $\subseteq$ *LHS*: Let $x \in S$ and assume that for all $y \in S$ there is a T^y such that $x \in \gamma(T^y)$ and $y \in T^y$. Thus $x \in \bigcap \gamma(T^y)$ and $S \subseteq \bigcup T^y$. From the former and the additivity of γ we get $x \in \gamma(\bigcup T^y)$, by (II), and now the latter and (I) yield $x \in \gamma(S)$, as desired. Q.E.D.

Proof of Lemma 2. (a) Let γ be $12n$-covering and satisfy (I). We show the claim for $\mathscr{P}_2(\gamma)$, which is identical with $\mathscr{P}(\gamma)$, by Lemma 1(c). Suppose for reductio that $x_1 <^2_\gamma x_2 <^2_\gamma \cdots <^2_\gamma x_n <^2_\gamma x_1$ for some $x_1, \ldots x_n \in X$. That is, $\gamma(\{x_i, x_{i+1}\}) = \{x_{i+1}\}$ with $+$ denoting addition modulo n. Now consider $\gamma(\{x_1, \ldots, x_n\}) \neq \emptyset$. Let $x_k \in \gamma(\{x_1, \ldots, x_n\})$. But $x_k \notin \gamma(\{x_k, x_k + 1\})$. This is in contradiction with (I). The rest of (a) is trivial.

(b) Let γ be 123-covering and satisfy (I) and (III). We show the claim for $\mathscr{P}_2(\gamma)$, which is identical with $\mathscr{P}(\gamma)$, by Lemma 1(c). Assume that $x \not\leq_{\gamma,2} y$ and $y \not\leq_{\gamma,2} z$. By definition of $\leq_{\gamma,2}$, this means that $y \notin \gamma(\{x, y\})$ and $z \notin \gamma(\{y, z\})$. Now consider $\gamma(\{x, y, z\})$. By (I), $y \notin \gamma(\{x, y, z\})$ and $z \notin \gamma(\{x, y, z\})$, so $\gamma(\{x, y, z\}) = \{x\}$ since $\gamma(\{x, y, z\})$ must be a non-empty subset of $\{x, y, z\}$. By (I), $x \in \gamma(\{x, z\})$, so $\gamma(\{x, y, z\}) \subseteq \gamma(\{x, z\})$. Hence, by (III), $\gamma(\{x, z\}) \subseteq \gamma(\{x, y, z\}) = \{x\}$, so $z \notin \gamma(\{x, z\})$, i.e. $x \not\leq_{\gamma,2} z$, as desired.

(c) Let γ be finitely additive and satisfy (IV), and let $x \leq_\gamma y$ and $y \leq_\gamma z$. This means that there is an $S \in \mathscr{X}$ such that $y \in \gamma(S)$ and $x \in S$ and a $T \in \mathscr{X}$ such that $z \in \gamma(T)$ and $y \in T$. Now consider $S \cup T$. Obviously, $x \in S \cup T$. In order to show that $x \leq_\gamma z$ it suffices to show that $z \in \gamma(S \cup T)$. By finite additivity, $S \cup T \in \mathscr{X}$. By $z \in \gamma(T)$ and (IV), it suffices to show that $\gamma(S \cup T) \cap T \neq \emptyset$. Suppose for reductio that $\gamma(S \cup T) \cap T = \emptyset$. Then, since $\emptyset \neq \gamma(S \cup T) \subseteq S \cup T$, $\gamma(S \cup T) \cap S \neq \emptyset$. So by (IV), $\gamma(S) \subseteq \gamma(S \cup T)$. So $y \in \gamma(S \cup T)$. But since also $y \in T$, $\gamma(S \cup T) \cap T \neq \emptyset$ after all, and we have a contradiction.

(Notice that an attempted proof of the transitivity of $\leq_{\gamma,2}$ would also need, apart from the 123-covering condition, (I) in order to come from $z \in \gamma(\{x, y, z\})$ to $z \in \gamma(\{x, z\})$, so we can rest content with (c).) Q.E.D.

Proof of Lemma 3. (a) Assume that $\mathscr{S}(\leq)$ is no selection function over $\mathscr{X}$. By the definition of $\mathscr{S}(\leq)$, this can only happen if some $S \in \mathscr{X}$ possesses no greatest element under $\leq$. Thus for every $x_i \in S$ there is an $x_{i+1} \in S$ such that $x_{i+1} \not\leq x_1$.

This, however, contradicts smoothness. That $\mathscr{S}(\preceq)$ satisfies (I) and (II) follows from Lemma 1(a).

(b) Let γ be relational with respect to some negatively transitive and negatively well-founded relation $\preceq$. (Or alternatively, let γ be subtractive and relational with respect to some negatively transitive and $\mathscr{X}$-smooth relation $\preceq$.) Let $S, S' \in \mathscr{X}$, $S \subseteq S'$ and $\gamma(S') \subseteq \gamma(S)$. We want to show that $\gamma(S) \subseteq \gamma(S')$. Suppose for reductio that this is not the case, i.e., that there is some x which is in $\gamma(S)$ but not in $\gamma(S')$. The latter means that there is some y_1 in S' such that $y_1 \not\preceq x$. As $x \in \gamma(S)$, $y_1 \in S' - S$. But because $\gamma(S') \subseteq \gamma(S) \subseteq S$, $y_1 \notin \gamma(S')$. So there is some $y_2 \in S'$ such that $y_2 \not\preceq y_1$. By negative transitivity, $y_2 \not\preceq x$. So by the same reasoning as before, $y_2 \in S' - S$ and $y_2 \notin \gamma(S')$. So there is some $y_3 \in S'$ such that $y_3 \not\preceq y_2$. By negative transitivity again, $y_3 \not\preceq x$, and the same reasoning can be continued again and again. What we get is an infinite chain $y_1, y_2, y_3, \ldots$ in $S' - S$ such that $\cdots \not\preceq y_3 \not\preceq y_2 \not\preceq y_1$. But this contradicts the negative well-foundedness of $\preceq$ (or the subtractivity of γ, which guarantees that $S' - S \in \mathscr{X}$, and smoothness of $\preceq$).

(c) Let γ be relational with respect to some transitive relation $\preceq$. Let $S, S' \in \mathscr{X}$, $S \subseteq S'$ and $x \in \gamma(S') \cap S$. By relationality, the latter conditions says that $y \preceq x$ for all $y \in S'$. Let $z \in \gamma(S)$. We have to show that $z \in \gamma(S')$, i.e., by relationality, that $y \preceq z$ for all $y \in S'$. But since $x \in S$ and $z \in \gamma(S)$, $x \preceq z$, so since $y \preceq x$ for all $y \in S'$, the desired conclusion follows from the transitivity of $\preceq$. Q.E.D.

Proof of Lemma 4. (a) Let γ be subtractive.

(I) implies (I′): Let $S, S', S \cup S' \in \mathscr{X}$ and $x \in \gamma(S \cup S')$. As $S, S' \subseteq S \cup S'$ and x is either in S or in S', we get $x \in \gamma(S) \cup \gamma(S')$, by (I), as desired. *(I′) implies (I)*: Let $S, S' \in \mathscr{X}$ and $S \subseteq S'$. By subtractivity, $S' - S \in \mathscr{X}$ as well. Using $S' = S \cup (S' - S)$ and (I′), we get $\gamma(S') \subseteq \gamma(S) \cup \gamma(S' - S)$. Intersecting both sides with S, then, we get $\gamma(S') \cap S \subseteq (\gamma(S) \cap S) \cup (\gamma(S' - S) \cap S) = \gamma(S)$, since $\gamma(S) \subseteq S$

(b) First we note that for finitely additive γ, (II′) is equivalent to

(II$_{fin}$) For all $\{S_i : i \in I, I \text{ finite}\} \subseteq \mathscr{X}$ such that $\bigcup\{S_i : i \in I\} \in \mathscr{X}$, $\bigcap\{\gamma(S_i) : i \in I\} \subseteq \gamma(\bigcup\{S_i : i \in I\})$

Now we simply modify the proof of Lemma 1(f) by taking only finitely many of the T^y's into consideration. This is possible because compactness can be applied to the inclusion $S \subseteq \bigcup T^y$ in that proof.

(c) Let γ satisfy (I).

(III) implies (III′): Let $S, S \cup S' \in \mathscr{X}$ and $\gamma(S \cup S') \cap S' = \emptyset$. From the latter condition and $\gamma(S \cup S') \subseteq S \cup S'$ it follows that $\gamma(S \cup S') \subseteq S$. Hence, by (I), $\gamma(S \cup S') = S \cap \gamma(S \cup S') \subseteq \gamma(S)$. Thus, by (III), $\gamma(S) \subseteq \gamma(S \cup S')$, as desired. *(III′) implies (III)*: Let $S, S' \in \mathscr{X}$, $S \subseteq S'$ and $\gamma(S') \subseteq \gamma(S)$. Since $S' = S \cup (S' - S) \in \mathscr{X}$ and $\gamma(S') \subseteq \gamma(S) \subseteq S$, it holds that $\gamma(S \cup (S' - S)) \cap (S' - S) = \emptyset$. By (III′), we get $\gamma(S) \subseteq \gamma(S')$, as desired.

(d) Immediate.

(e) That (I&II) implies (I) is immediate if we put $I = \{i\}$ and $S_i = S'$, and that (I&II) implies (II) is immediate if we put $S = \bigcup S_i$ and observe that $\gamma(S_i) \subseteq S$.

In order to see that (I) and (II) taken together imply (I&II), we note that $\bigcup\{S_i : i \in I\} \in \mathscr{X}$, by additivity. So (II) gives us $S \cap \bigcap \gamma(S_i) \subseteq S \cap \gamma(\bigcup S_i)$, and (I) gives us $S \cap \gamma(\bigcup S_i) \subseteq \gamma(S)$, whenever $S \subseteq \bigcup S_i$. So in this case $S \cap \bigcap \gamma(S_i) \subseteq \gamma(S)$, as desired.

(f) Similar to (e). Q.E.D.

Re section "Representation Theorems for Contraction Functions": Example of a relational partial meet contraction function which does not satisfy (÷8r).

Consider a propositional language L with denumerably many atoms $p_1, p_2, p_3, \ldots$ and q and r. Let $A = Cn(p_1, p_2, p_3, \ldots, q, r)$. Let W_i, W_i^q, W_i^r, for $i = 1, 2, 3, 4, \ldots$, be the maximally consistent set of propositions defined by

$p_i \in W_i, W_i^q, W_i^r$, for every i,
$\neg p_j \in W_i, W_i^q, W_i^r$, for every i and every $j \neq i$,
$q \in W_i^q$ and $\neg q \in W_i, W_i^r$, for every i,
$r \in W_i^r$ and $\neg r \in W_i, W_i^q$, for every i.

Let $\mathbf{W}^{qr} = \bigcup\{\{W_i, W_i^q, W_i^r\} : i = 1, 2, 3, \ldots\}$. Remember that $V_A = \mathbf{W} - [\![A]\!]$. Let the relation $<_W$ over V_A be defined as follows:

$W <_W W'$ for every $W \in \mathbf{W}^{qr}$ and $W' \in V_A - \mathbf{W}^{qr}$
$W_j^q <_W W_{j+1}$, $W_j^q <_W W_{j+1}^q$, and $W_j^q <_W W_{j+1}^r$, for every $j \in \{0, 2, 4, 6, \ldots\}$
$W_k^r <_W W_{k+1}$, $W_k^r <_W W_{k+1}^q$, and $W_k^r <_W W_{k+1}^r$, for every $k \in \{1, 3, 5, \ldots\}$
These are all pairs standing in the relation $<_W$ (see Fig. 16.3).

Then define, for the purposes of this example, $M \leq M'$ iff $M' \not< M$ iff $\mathscr{W}(M) \not<_W \mathscr{W}(M')$. Then clearly, the $\leq$-greatest elements of $A \perp x$ correspond to the $<$-minimal elements of $[\![\neg x]\!]$. Note that $\leq$ is connected, acyclic and conversely well-founded (no infinite ascending chains), but not negatively transitive. Let $\dot{-} = \mathscr{C}(\mathscr{S}(\leq))$ be the relational partial meet contraction determined by $\leq$.

Now observe that the greatest elements of $A \perp (q \wedge r) = A \perp q \cup A \perp r$ are exactly $A \cap W_0$, $A \cap W_0^q$, and $A \cap W_0^r$. Since p_0 is an element of all these sets, we have $p_0 \in A \dot{-} (q \wedge r)$.

The greatest elements of $A \perp q$ are $A \cap W_k$ and $A \cap W_k^r$ for every $k \in \{0, 1, 3, 5, \ldots\}$. Obviously, $q \rightarrow x \in A \dot{-} q$ for every $x \in A$. Furthermore, $\neg p_n \vee \neg p_m \in A \dot{-} q$ for every $n \neq m$. Informally speaking, $A \dot{-} q$ contains the information that exactly one proposition p_k, for $k \in \{0, 1, 3, 5, \ldots\}$, is true. But this cannot be expressed in our first-order propositional language. Note in particular that $A \dot{-} q$ does not contain any finite disjunction of p_i's.

The greatest elements of $A \perp r$ are $A \cap W_j$ and $A \cap W_j^q$ for every $j \in \{0, 2, 4, 6, \ldots\}$. The information contained in $A \dot{-} r$ is analogous with that in $A \dot{-} q$, with the even positive integers playing the role of the odd ones in the previous case.

We find that $A \dot{-} q$ and $A \dot{-} r$ do not even jointly imply p_0. For consider the maximally consistent set W^* containing $\neg p_i$ for every $i \geq 0$, together with $\neg q$ and $\neg r$. Considering what we have just said about $A \dot{-} q$ and $A \dot{-} r$, it becomes clear that W^* contains both of these contracted theories, as well as $\neg p_0$.

Since $p_0 \in A \dot{-} (q \wedge r)$ but $p_0 \notin Cn(A \dot{-} q \cup A \dot{-} r)$, we have a violation of (÷8r).

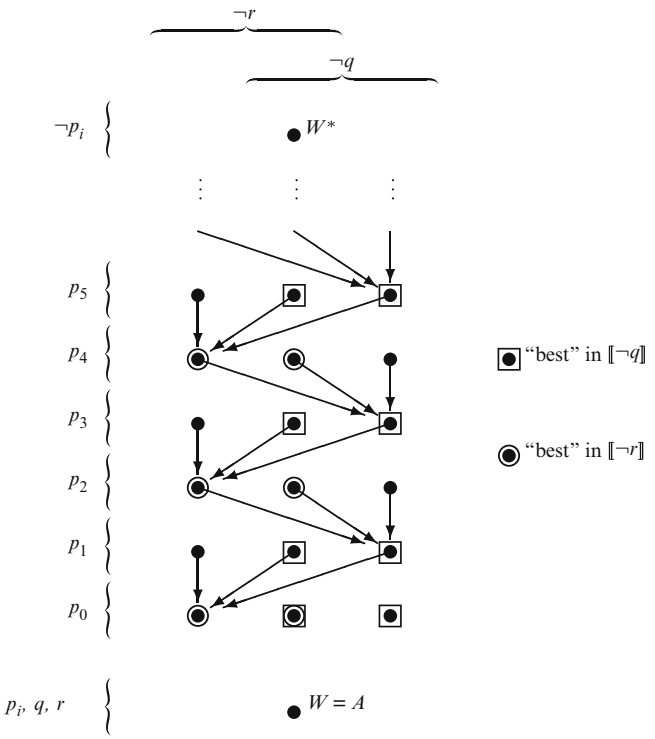


Fig. 16.3 Relational partial meet contraction not satisfying (÷8r)

Proof of Lemma 9. We show that for every $M \in A\perp x$,

(i) M is maximal in $A\perp x$ under $\lll$
is equivalent to
(ii) $N \subseteq M$ for some N which is maximal in $B\perp x$ under $\lll$.

(i) implies (ii): (i) means that M is such that for no M' and i it holds that $M \lll_i M'$. Now take $N = M \cap B$. We have to show that N is maximal in $B\perp x$. First, it is clear that $N \nvdash x$. Next, we verify that N is a maximal subset of B which does not imply x. Suppose there is a $y \in B - N$ such that $N \cup \{y\} \nvdash x$, but then, with $y \in B_k$, $N_{\geq k} \cup \{y\} \nvdash x$. So there would be a superset M' of $N_{\geq k} \cup \{y\}$ in $A\perp x$ for which $M \lll_{\geq k} M'$, contradicting (i). Finally, we have to show that there is no $N' \in B\perp x$ which is better than N under $\lll$. But by the same argument as just used, it is impossible that there be some $y \in B_k$ such that $y \in N' - N$ and $N \lll_k N'$.

(ii) implies (i): Suppose for reductio that (ii) but not (i). By the latter, there is an M' such that $M \lll M'$. But then $M \cap B \lll M' \cap B$. Take, on the other hand, the N from (ii). Since $N \subseteq M \cap B$, we have $N \lll M' \cap B$. But since $M' \cap B \nvdash x$, N cannot be maximal in $B\perp x$, contradicting (ii). Q.E.D.

Re section "The Logic of Prioritized Base Contractions in the Finite Case": Illustration of the counterexample to (⊢8r) in straightforward prioritized base contractions and its solution by the official definition. Let $B = \{(p \wedge q) \vee (p \wedge r) \vee (q \wedge r),\ r \rightarrow (p \leftrightarrow q),\ q \wedge (p \leftrightarrow r),\ p \wedge r\}, A = Cn(B), A \dot{-}' x = \bigcap \{Cn(N) : N \in \gamma(B \perp x)\}$ and $A \dot{-} x = \bigcap \{Cn(N \cup \{x \rightarrow a\}) : N \in \gamma(B \perp x)\}$. See Fig. 16.4

Re section "The Logic of Prioritized Base Contractions in the Finite Case": Example illustrating how prioritized belief bases can be replaced by simple ones.

Consider the propositional language L over the two atoms p and q. Let the belief base contain the propositions p and $p \rightarrow q$. We contrast the prioritizations $\langle \{p \rightarrow q\}, \{p\} \rangle$ and $\langle \{p\}, \{p \rightarrow q\} \rangle$. In the former case, $B = \mathscr{B}(\mathscr{P}(\langle \{p \rightarrow q\}, \{p\} \rangle)) = \{p \wedge q, p, p \vee q, q \rightarrow p\}$, while in the latter case, $B' = \mathscr{B}(\mathscr{P}(\langle \{p\}, \{p \rightarrow q\} \rangle)) = \{p \wedge q, q, p \leftrightarrow q, \top\}$.

In Figs. 16.5 and 16.6, "$\binom{x}{y}{z}$" should be read as "$A \dot{-} x = Cn(y)$ according to prioritized base contraction, while $A \dot{-} x = Cn(z)$ according to the corresponding simple base contraction". In the last two lines, the left argument of '$\wedge$' comes from straightforward base contraction and the right argument is the recovery-guaranteeing appendage used in our official definition of prioritized base contractions.

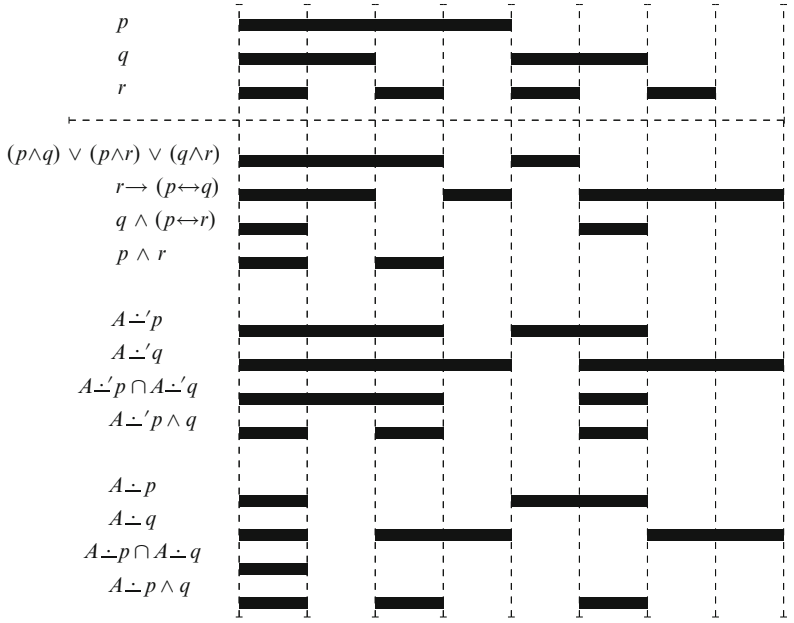


Fig. 16.4 Prioritized base contraction $\dot{-}'$ with simple base violating (⊢8r), and strengthened contraction $\dot{-}$ satisfying (⊢8r)

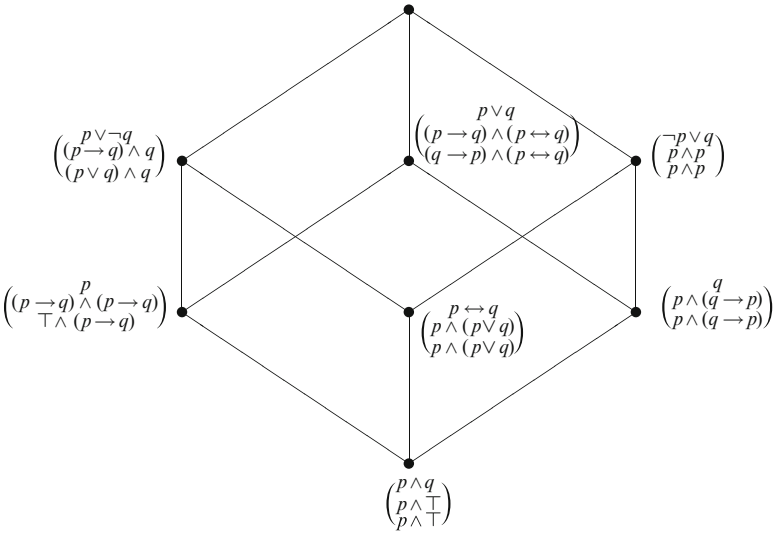


Fig. 16.5 Prioritized base contraction of a two-element base, with p having priority over $p \to q$

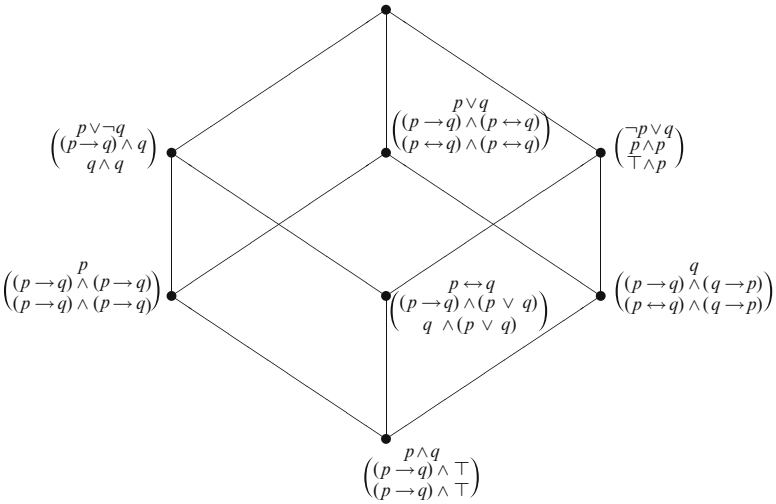


Fig. 16.6 Prioritized base contraction of a two-element base, with $p \to q$ having priority over p

Remarks. Prioritized and corresponding simple base contractions indeed lead to the same results—as they should, according to the proof of Theorem 7. But note that the recovery-guaranteeing appendage is essential in quite a few cases. As for the contraction function $\dot{-}$, the differences in prioritization are effective only in the case of $A\dot{-}(p \wedge q)$, $A\dot{-}q$ and $A\dot{-}(p \leftrightarrow q)$.

References

Alchourrón, C., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction functions and their associated revision functions. *Journal of Symbolic Logic, 50*, 510–530.

Alchourrón, C., & Makinson, D. (1986). Maps between some different kinds of contraction function: The finite case. *Studia Logica, 45*, 187–198.

Arrow, K. J. (1959). Rational choice functions and orderings. *Economica*, N.S., *26*, 121–127.

Chernoff, H. (1954). Rational selection of decision functions. *Econometrica, 22*, 422–443.

de Condorcet, N. (1785). Essai sur l'application de l'analyse á la probabilité des décisions rendues á la pluralité des voix. Paris: Imprimerie Royale. Reprint Cambridge: Cambridge University Press (2014).

Fuhrmann, A., & Morreau, M. (Eds.) (1991). *The logic of theory change* (Lecture notes in computer science, Vol. 465). Berlin: Springer.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge: Bradford Books/MIT.

Gärdenfors, P. (Ed.) (1992). *Belief revision*. Cambridge: Cambridge University Press.

Gärdenfors, P., & Makinson, D. (1988). Revisions of knowledge systems using epistemic entrenchment. In M. Vardi (Ed.), *Theoretical aspects of reasoning about knowledge* (pp. 83–95). Los Altos: Morgan Kaufmann.

Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic, 17*, 157–170.

Hansson, S. O. (1992). Similarity semantics and minimal changes of belief. *Erkenntnis, 37*, 401–429.

Herzberger, H. G. (1973). Ordinal preference and rational choice. *Econometrica, 41*, 187–237.

Katsuno, H., & Mendelzon, A. O. (1991). Propositional knowledge base revision and minimal change. *Artificial Intelligence, 52*, 263–294.

Kraus, S., Lehmann, D., & Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence, 44*, 167–207.

Lehmann, D., & Magidor, M. (1992). What does a conditional knowledge base entail? *Artificial Intelligence, 55*, 1–60.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1981). Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic, 10*, 217–234.

Lindström, S. (1991). A semantic approach to nonmonotonic reasoning: Inference and choice. University of Uppsala, April 1991 (manuscript).

Makinson, D., & Gärdenfors, P. (1991). Relations between the logic of theory change and nonmonotonic logic. In Fuhrmann & Morreau (1991) (pp. 185–205).

Nebel, B. (1989). A knowledge level analysis of belief revision. In R. Brachman, H. Levesque, & R. Reiter (Eds.), *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning* (pp. 301–311). San Mateo: Morgan Kaufmann.

Nebel, B. (1992). Syntax-based approaches to belief revision. In Gärdenfors (1992) (pp. 52–88).

Rott, H. (1991). Two methods of constructing contractions and revisions of knowledge systems. *Journal of Philosophical Logic, 20*, 149–173.

Rott, H. (1992a). On the logic of theory change: More maps between different kinds of contraction function. In Gärdenfors (1992) (pp. 122–141).

Rott, H. (1992b). Preferential belief change using generalized epistemic entrenchment. *Journal of Logic, Language and Information, 1*, 45–78.

Rott, H. (2003). Basic entrenchment. *Studia Logica*, *73*, 257–280.

Samuelson, P. A. (1950). The problem of integrability in utility theory. *Economica*, N.S., *17*, 355–381.

Sen, A. K. (1982) *Choice, welfare and measurement*. Oxford: Blackwell.

Suzumura, K. (1983) *Rational choice, collective decisions, and social welfare*. Cambridge: Cambridge University Press.

Uzawa, H. (1956). Note on preference and axioms of choice. *Annals of the Institute of Statistical Mathematics, 8*, 35–40.

Chapter 17
A Survey of Ranking Theory

Wolfgang Spohn

Introduction

Epistemology is concerned with the fundamental laws of thought, belief, or judgment. It may inquire the fundamental relations among the objects or contents of thought and belief, i.e., among propositions or sentences. Then we enter the vast realm of formal logic. Or it may inquire the activity of judging or the attitude of believing itself. Often, we talk as if this would be a yes or no affair. From time immemorial, though, we know that judgment is firm or less than firm, that belief is a matter of degree. This insight opens another vast realm of formal epistemology.

Logic received firm foundations already in ancient philosophy. It took much longer, though, until the ideas concerning the forms of (degrees of) belief acquired more definite shape. Despite remarkable predecessors in Indian, Greek, Arabic, and medieval philosophy, the issue seemed to seriously enter the agenda of intellectual history only in the sixteenth century with the beginning of modern philosophy. Cohen (1980) introduced the handy, though somewhat tendentious opposition between Baconian and Pascalian probability. This suggests that the opposition was already perceivable with the work of Francis Bacon (1561–1626) and Blaise Pascal (1623–1662). In fact, philosophers were struggling to find the right mould. In that struggle, Pascalian probability, which *is* probability *simpliciter*, was the first to take a clear and definite shape, viz. in the middle of seventeenth century (cf. Hacking 1975), and since then it advanced triumphantly. The extent to which it interweaves with our cognitive enterprise has become nearly total (cf. the marvelous collection of Krüger et al. 1987). There certainly were alternative ideas. However, probability theory was always far ahead; indeed, the distance ever increased. The winner takes it all!

W. Spohn (✉)
Fachbereich Philosophie, Universität Konstanz, 78457 Konstanz, Germany
e-mail: wolfgang.spohn@uni-konstanz.de

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_17

I use 'Baconian probability' as a collective term for the alternative ideas. This is legitimate since there are strong family resemblances among the alternatives. Cohen has chosen an apt term since it gives historical depth to ideas that can be traced back at least to Bacon (1620) and his powerful description of 'the method of lawful induction'. Jacob Bernoulli and Johann Heinrich Lambert struggled with a non-additive kind of probability. When Joseph Butler and David Hume spoke of probability, they often seemed to have something else or more general in mind than our precise explication. In contrast to the German Fries school British nineteenth century's philosophers like John Herschel, William Whewell, and John Stuart Mill elaborated non-probabilistic methods of inductive inference. And so forth.[1]

Still, one might call this an underground movement. The case of alternative forms of belief became a distinct hearing only in the second half of the twentieth century. On the one hand, there were scattered attempts like the 'functions of potential surprise' of Shackle (1949), heavily used and propagated in the epistemology of Isaac Levi since his (1967), Rescher's (1964) account of hypothetical reasoning, further developed in his (1976) into an account of plausible reasoning, or Cohen's (1970) account of induction which he developed in his (1977) under the label 'Non-Pascalian probability', later on called 'Baconian'. On the other hand, one should think that modern philosophy of science with its deep interest in theory confirmation and theory change produced alternatives as well. Indeed, Popper's hypothetical-deductive method proceeded non-probabilistically, and Hempel (1945) started a vigorous search for a qualitative confirmation theory. However, the former became popular rather among scientists than among philosophers, and the latter petered out after 25 years, at least temporarily.

I perceive all this rather as a prelude, preparing the grounds. The outburst came only in the mid 70s, with strong help from philosophers, but heavily driven by the needs of Artificial Intelligence. Not only deductive, but also inductive reasoning had to be implemented in the computer, probabilities appeared intractable[2], and thus a host of alternative models were invented: a plurality of default logics, non-monotonic logics and defeasible reasonings, fuzzy logic as developed by Zadeh (1975, 1978), possibility theory as initiated by Zadeh (1978) and developed by Dubois and Prade (1988), the Dempster-Shafer belief functions originating from Dempster (1967, 1968), but essentially generalized by Shafer (1976), AGM belief revision theory (cf. Gärdenfors 1988), a philosophical contribution with great success in the AI market, Pollock's theory of defeasible reasoning (summarized in Pollock 1995), and so forth. The field has become rich and complex. There are attempts of unification like Halpern (2003) and huge handbooks like Gabbay et al. (1994). One hardly sees the wood for trees. It seems that what had been forgotten for centuries had to be made good for within decades.

[1] This is not the place for a historical account. See, e.g., Cohen (1980) and Shafer (1978) for some details.

[2] Only Pearl (1988) showed how to systematically deal with probabilities without exponential computational explosion.

Ranking theory, first presented in Spohn (1983, 1988)[3], belongs to this field as well. Since its development, by me and others, is scattered in a number of papers, one goal of the present paper is to present an accessible survey of the present state of ranking theory.[4] This survey will emphasize the philosophical applications, thus reflecting my bias towards philosophy. My other goal is justificatory. Of course, I am not so blinded to claim that ranking theory would be *the* adequate account of Baconian probability. As I said, 'Baconian probability' stands for a collection of ideas united by family resemblances; and I shall note some of the central resemblances in the course of the paper. However, there is a multitude of epistemological purposes to serve, and it is entirely implausible that there is one account to serve all. Hence, postulating a reign of probability is silly, and postulating a duumvirate of probability and something else is so, too. Still, I am not disposed to see ranking theory as just one offer among many. On many scores, ranking theory seems to me to be superior to rival accounts, the central score being the notion of *conditional* ranks. I shall explain what these scores are, thus trying to establish ranking theory as one particularly useful account of the laws of thought.

The plan of the paper is simple. In the five subsections of section "The theory", pp. 305ff, I shall outline the main aspects of ranking theory. This central section will take some time. I expect the reader to get impatient meanwhile; you will get the strong impression that I am not presenting an alternative to (Pascalian) probability, as the label 'Baconian' suggests, but simply probability itself in a different disguise. This is indeed one way to view ranking theory, and a way, I think, to understand its virtues. However, the complex relation between probability and ranking theory, though suggested at many earlier points, will be systematically discussed only in section "Ranks and probabilities", pp. 328ff. The section "Further comparisons", pp. 335ff, will finally compare ranking theory to some other accounts of Baconian probability.

The Theory

Basics

We have to start with fixing the objects of the cognitive attitudes we are going to describe. This is a philosophically highly contested issue, but here we shall stay conventional without discussion. These objects are pure contents, i.e., propositions.

[3]There I called its objects ordinal conditional functions. Goldszmidt and Pearl (1996) started calling them ranking functions, a usage I happily adapted.

[4]In the meantime, my comprehensive book on ranking theory (Spohn 2012) has appeared. This paper may also serve as an introduction to this book. Reversely, various topics, which are only touched here and then referred back to older papers of mine, are developed in this book in a better and more comprehensive way.

To be a bit more explicit: We assume a non-empty set W of mutually exclusive and jointly exhaustive possible worlds or *possibilities*, as I prefer to say, for avoiding the grand associations of the term 'world' and for allowing to deal with *de se* attitudes and related phenomena (where doxastic alternatives are considered to be centered worlds rather than worlds). And we assume an algebra $\mathcal{A}$ of subsets of W, which we call *propositions*. All the functions we shall consider for representing doxastic attitudes will be functions defined on that algebra $\mathcal{A}$.

Thereby, we have made the philosophically consequential decision of treating doxastic attitudes as intensional. That is, when we consider sentences such as "*a* believes (with degree *r*) that *p*", then the clause *p* is substitutable salva veritate by any clause *q* expressing the same proposition and in particular by any logically equivalent clause *q*. This is so because by taking propositions as objects of belief we have decided that the truth value of such a belief sentence depends only on the proposition expressed by *p* and not on the particular way of expressing that proposition. The worries provoked by this decision are not our issue.

The basic notion of ranking theory is very simple:

Definition 17.1 Let $\mathcal{A}$ be an algebra over W. Then κ is a *negative ranking function*[5] for $\mathcal{A}$ iff κ is a function from $\mathcal{A}$ into $\mathbf{R}^* = \mathbf{R}^+ \cup \{\infty\}$ (i.e., into the set of non-negative reals plus infinity) such that for all $A, B \in \mathcal{A}$:

$$\kappa(W) = 0 \text{ and } \kappa(\varnothing) = \infty, \tag{17.1}$$

$$\kappa(A \cup B) = \min\{\kappa(A), \kappa(B)\} \qquad [\textit{the law of disjunction (for negative ranks)}]\,. \tag{17.2}$$

$\kappa(A)$ is called the (*negative*) *rank* of A.

It immediately follows for each $A \in \mathcal{A}$:

$$\text{either } \kappa(A) = 0 \text{ or } \kappa(\overline{A}) = 0 \text{ or both} \qquad [\textit{the law of negation}]\,. \tag{17.3}$$

A negative ranking function κ, this is the standard interpretation, expresses a *grading of disbelief* (and thus something negative, hence the qualification). If $\kappa(A) = 0$, A is not disbelieved at all; if $\kappa(A) > 0$, A is disbelieved to some positive degree. Belief in A is the same as disbelief in $\overline{A}$; hence, A is *believed* in κ iff $\kappa(\overline{A}) > 0$. This entails (via the law of negation), but is not equivalent to $\kappa(A) = 0$. The latter is compatible also with $\kappa(\overline{A}) = 0$, in which case κ is neutral or unopinionated concerning A. We shall soon see the advantage of explaining belief in this indirect way via disbelief.

A little example may be instructive. Let us look at Tweetie of which default logic is very fond. Tweetie has, or fails to have, each of the three properties: being a bird

[5]For systematic reasons I am slightly rearranging my terminology from earlier papers. I would be happy if the present terminology became the official one.

(B), being a penguin (P), and being able to fly (F). This makes for eight possibilities. Suppose you have no idea what Tweetie is, for all you know it might even be a car. Then your ranking function may be the following one, for instance:[6]

κ	$B\&\overline{P}$	$B\ \&\ P$	$\overline{B}\&\overline{P}$	$\overline{B}\&P$
F	0	4	0	11
$\overline{F}$	2	1	0	8

In this case, the strongest proposition you believe is that Tweetie is *either* no penguin and no bird $\left(\overline{B}\&\overline{P}\right)$ *or* a flying bird and no penguin ($F\ \&\ B\ \&\ \overline{P}$). Hence, you neither believe that Tweetie is a bird (B) nor that it is not a bird ($\overline{B}$). You are also neutral concerning its ability to fly. But you believe, for instance: if Tweetie is a bird, it is not a penguin and can fly $\left(B \rightarrow \overline{P}\&F\right)$; and if Tweetie is not a bird, it is not a penguin $\left(\overline{B} \rightarrow \overline{P}\right)$ – each if-then taken as material implication. In this sense you also believe: if Tweetie is a penguin, it can fly ($P \rightarrow F$); and if Tweetie is a penguin, it cannot fly $\left(P \rightarrow \overline{F}\right)$ – but only because you believe that it is not a penguin in the first place; you simply do not reckon with its being a penguin. If we understand the if-then differently, as we shall do later on, the picture changes. The larger ranks in the last column indicate that you strongly disbelieve that penguins are not birds. And so we may discover even more features of this example.

What I have explained so far makes clear that we have already reached the first fundamental aim ranking functions are designed for: the *representation of belief*. Indeed, we may define $\mathcal{B}_\kappa = \{A \mid \kappa(\overline{A}) > 0\}$ to be the *belief set* associated with the ranking function κ. This belief set is finitely *consistent* in the sense that whenever $A_1, \ldots, A_n \in \mathcal{B}_\kappa$, then $A_1 \cap \ldots \cap A_n \neq \varnothing$; this is an immediate consequence of the law of negation. And it is finitely *deductively closed* in the sense that whenever $A_1, \ldots, A_n \in \mathcal{B}_\kappa$ and $A_1 \cap \ldots \cap A_n \subseteq B \in \mathcal{A}$, then $B \in \mathcal{B}_\kappa$; this is an immediate consequence of the law of disjunction. Thus, belief sets just have the properties they are normally assumed to have. (The finiteness qualification is a little cause for worry that will be addressed soon.)

There is a big argument about the rationality postulates of consistency and deductive closure; we should not enter it here. Let me only say that I am disappointed by all the attempts I have seen to weaken these postulates. And let me point out that the issue was essentially decided at the outset when we assumed belief to operate on propositions or truth-conditions or sets of possibilities. With these assumptions we ignore the relation between propositions and their sentential expressions or modes of presentation; and it is this relation where all the problems hide.

[6] I am choosing the ranks in an arbitrary, though intuitively plausible way (just as I would have to arbitrarily choose plausible subjective probabilities, if the example were a probabilistic one). The question how ranks may be measured will be taken up in section "The dynamics of belief and the measurement of belief", pp. 316ff.

When saying that ranking functions represent belief I do not want to further qualify this. One finds various notions in the literature, full beliefs, strong beliefs, weak beliefs, one finds a distinction of acceptance and belief, etc. In my view, these notions and distinctions do not respond to any settled intuitions; they are rather induced by various theoretical accounts. Intuitively, there is only one perhaps not very clear, but certainly not clearly divisible phenomenon which I exchangeably call believing, accepting, taking to be true, etc.

However, if the representation of belief were our only aim, belief sets or their logical counterparts as developed in doxastic logic (see already Hintikka 1962) would have been good enough. What then is the purpose of the ranks or degrees? Just to give another account of the intuitively felt fact that belief is graded? But what guides such accounts? Why should degrees of belief behave like ranks as defined? Intuitions by themselves are not clear enough to provide this guidance. Worse still, intuitions are usually tainted by theory; they do not constitute a neutral arbiter. Indeed, problems already start with the intuitive conflict between representing belief and representing degrees of belief. By talking of belief *simpliciter*, as I have just insisted, I seem to talk of *ungraded* belief.

The only principled guidance we can get is a theoretical one. The degrees must serve a clear theoretical purpose and this purpose must be shown to entail their behavior. For me, the theoretical purpose of ranks is unambiguous; this is why I invented them. It is the *representation of the dynamics of belief*; that is the second fundamental aim we pursue. How this aim is reached and why it can be reached in no other way will unfold in the course of this section. This point is essential; as we shall see, it distinguishes ranking theory from all similarly looking accounts, and it grounds its superiority.

For the moment, though, let us look at a number of variants of Definition 17.1. Above I mentioned the finiteness restriction of consistency and deductive closure. I have always rejected this restriction. An inconsistency is irrational and to be avoided, be it finitely or infinitely generated. Or, equivalently, if I take to be true a number of propositions, I take their conjunction to be true as well, even if the number is infinite. If we accept this, we arrive at a somewhat stronger notion:

Definition 17.2 Let $\mathcal{A}$ be a complete algebra over W (closed also under infinite Boolean operations). Then κ is a *complete negative ranking function* for $\mathcal{A}$ iff κ is a function from W into $\mathbf{N}^+ = \mathbf{N} \cup \{\infty\}$ (i.e., into the set of non-negative integers plus infinity) such that $\kappa^{-1}(0) \neq \emptyset$ and and $\kappa^{-1}(n) \in \mathcal{A}$ for each $n \in \mathbf{N}^+$. κ is extended to propositions by defining $\kappa(\emptyset) = \infty$ and $\kappa(A) = \min\{\kappa(w) \mid w \in A\}$ for each non-empty $A \in \mathcal{A}$.

Obviously, the propositional function satisfies the laws of negation and disjunction. Moreover, we have for any $\mathcal{B} \subseteq \mathcal{A}$:

$$\kappa(\cup\mathcal{B}) = \min\{\kappa(B) \mid B \in \mathcal{B}\} \qquad [\textit{the law of infinite disjunction}]. \tag{17.4}$$

Due to completeness, we could start in Definition 17.2 with the point function and then define the set function as specified. Equivalently, we could have defined the set functions by the conditions (17.1) and (17.4) and then reduce the set function to a point function. Henceforth I shall not distinguish between the point and the set function. Note, though, that without completeness the existence of an underlying point function is not guaranteed; the relation between point and set function in this case is completely cleared up in Huber (2006).

Why are complete ranking functions confined to integers? The reason is condition (17.4). It entails that any infinite set of ranks has a minimum and hence that the range of a complete ranking function is well-ordered. Hence, the natural numbers are a natural choice. In my first publications (1983) and (1988) I allowed for more generality and assumed an arbitrary set of ordinal numbers as the range of a ranking function. However, since we want to calculate with ranks, this meant to engage into ordinal arithmetic, which is awkward. Therefore I later confined myself to complete ranking functions as defined above.

The issue about condition (17.4) was first raised by Lewis (1973, sect. 1.4) where he introduced the so-called Limit Assumption in relation to his semantics of counterfactuals. Endorsing (17.4), as I do, is tantamount to endorsing the Limit Assumption. Lewis finds reason against it, though it does not affect the *logic* of counterfactuals. From a semantic point of view, I do not understand his reason. He requests us to counterfactually suppose that a certain line is longer than an inch and asks how long it would or might be. He argues in effect that for each $\varepsilon > 0$ we should accept as true: "If the line would be longer than 1 inch, it would not be longer than $1 + \varepsilon$ inches." This strikes me as blatantly inconsistent, even if we cannot derive a contradiction in counterfactual logic (due to its ω–incompleteness). Therefore, I am accepting the Limit Assumption and, correspondingly, the law of infinite disjunction. This means in particular that in that law the minimum must not be weakened to the infimum.

Though I prefer complete ranking functions for the reasons given, the issue will have no further relevance here. In particular, if we assume the algebra of propositions to be finite, each ranking function is complete, and the issue does not arise. In the sequel, you can add or delete completeness as you wish.

Let me add another observation apparently of a technical nature. It is that we can mix ranking functions in order to form a new ranking function. This is the content of

Definition 17.3 Let Λ be a non-empty set of negative ranking functions for an algebra $\mathcal{A}$ of propositions, and let ρ be a complete negative ranking function over Λ. Then κ defined by

$$\kappa(A) = \min\{\lambda(A) + \rho(\lambda) \mid \lambda \in \Lambda\} \text{ for all } A \in \mathcal{A} \tag{17.5}$$

is obviously a negative ranking function for $\mathcal{A}$ as well and is called the *mixture of* Λ *by* ρ.

It is nice that such mixtures make formal sense. However, we shall see in the course of this paper that the point is more than a technical one; such mixtures will acquire deep philosophical importance later on.

So far, (degree of) disbelief was our basic notion. Was this necessary? Certainly not. We might just as well express things in positive terms:

Definition 17.4 Let $\mathcal{A}$ be an algebra over W. Then π is a *positive ranking function* for $\mathcal{A}$ iff π is a function from $\mathcal{A}$ into $\mathbf{R}^*$ such that for all $A, B \in \mathcal{A}$:

$$\pi(\varnothing) = 0 \text{ and } \pi(W) = \infty, \tag{17.6}$$

$$\pi(A \cap B) = \min\{\pi(A), \pi(B)\} \quad [\textit{the law of conjunction for positive ranks}]. \tag{17.7}$$

Positive ranks express *degrees of belief*. $\pi(A) > 0$ says that A is believed (to some positive degree), and $\pi(A) = 0$ says that A is not believed. Obviously, positive ranks are the dual to negative ranks; if $\pi(A) = \kappa(\overline{A})$ for all $A \in \mathcal{A}$, then π is a positive function iff κ is a negative ranking function.

Positive ranking functions seem distinctly more natural. Why do I still prefer the negative version? A superficial reason is that we have seen complete negative ranking functions to be reducible to point functions, whereas it would obviously be ill-conceived to try the same for the positive version. This, however, is only indicative of the main reason. Despite appearances, we shall soon see that negative ranks behave very much like probabilities. In fact, this parallel will serve as our compass for a host of exciting observations. (For instance, in the finite case probability measures can also be reduced to point functions.) If we were thinking in positive terms, this parallel would remain concealed.

There is a further notion that may appear even more natural:

Definition 17.5 Let $\mathcal{A}$ be an algebra over W. Then τ is a *two-sided ranking function*[7] for $\mathcal{A}$ iff τ is a function from $\mathcal{A}$ into $\mathbf{R} \cup \{-\infty, \infty\}$ such that there is a negative ranking function κ and its positive counterpart π for which for all $A \in \mathcal{A}$:

$$\tau(A) = \kappa(\overline{A}) - \kappa(A) = \pi(A) - \kappa(A).$$

Obviously, we have $\tau(A) > 0$, < 0, or $= 0$ according to whether A is believed, disbelieved, or neither. In this way, the belief values of all propositions are expressed in a single function. Moreover, we have the appealing law that $\tau(\overline{A}) = -\tau(A)$. For some purposes this is a useful notion that I shall readily employ. However, its formal behavior is awkward. Its direct axiomatic characterization would have been cumbersome, and its simplest definition consisted in its reduction to the other notions.

[7] In earlier papers I called this a belief function, obviously an unhappy term which has too many different uses. This is one reason fort the mild terminological reform proposed in this paper.

Still, this notion suggests an interpretational degree of freedom so far unnoticed.[8] We might ask: Why does the range of belief extend over all the positive reals in a two-sided ranking function and the range of disbelief over all the negative reals, whereas neutrality shrinks to rank 0? This looks unfair. Why may unopinionatedness not occupy a much broader range? Indeed, why not? We might just as well distinguish some positive rank or real z and define the closed interval $[-z, z]$ as the range of neutrality. Then $\tau(A) > z$ expresses belief in A and $\tau(A) < -z$ disbelief in A. This is a viable interpretation; in particular, consistency and deductive closure of belief sets would be preserved. However, 0 would still be a distinguished rank in this interpretation; it marks *central* neutrality, as it were, since it is the only rank x for which we may have $\tau(A) = \tau(\overline{A}) = x$.

The interpretational freedom appears quite natural. After all, the notion of belief is certainly vague and can be taken more or less strict. We can do justice to this vagueness with the help of the parameter z. The crucial point, though, is that we always get the formal structure of belief we want to get, however we fix that parameter. The principal lesson of this observation is, hence, that it is not the notion of belief which is of basic importance; it is rather the formal structure of ranks. The study of belief *is* the study of *that* structure. Still, it would be fatal to simply give up talking of belief in favor of ranks. Ranks express beliefs, even if there is interpretational freedom. Hence, it is of paramount importance to maintain the intuitive connection. In the sequel, I shall stick to my standard interpretation and equate belief in A with $\tau(A) > 0$, even though this is a matter of decision.

Let us pause for a moment and take a brief look back. What I have told so far probably sounds familiar. One has quite often seen all this, in this or a similar form – where the similar form may also be a comparative one: as long as only the ordering and not the numerical properties of the degrees of belief are relevant, a ranking function may also be interpreted as a weak ordering of propositions according to their plausibility, entrenchment, credibility, etc. Often things are cast in negative terms, as I primarily do, and often in positive terms. In particular, the law of negation securing consistency and the law of disjunction somehow generalizing deductive closure (we still have to look at the point more thoroughly) or their positive counterparts are pervasive. If one wants to distinguish a common core in that ill-defined family of Baconian probability, it is perhaps just these two laws.

So, why invent a new name, 'ranks', for familiar stuff? The reason lies in the second fundamental aim associated with ranking functions: to account for the dynamics of belief. This aim has been little pursued under the label of Baconian probability, but it is our central topic for the rest of this section. Indeed, everything stands and falls with our notion of conditional ranks; it is the distinctive mark of ranking theory. Here it is:

Definition 17.6 Let κ be a negative ranking function for $\mathcal{A}$ and $\kappa(A) < \infty$. Then the *conditional rank* of $B \in \mathcal{A}$ *given* A is defined as $\kappa(B \mid A) = \kappa(A \cap B) - \kappa(A)$.

[8] I am grateful to Matthias Hild for making this point clear to me.

The function $\kappa_A : B \mapsto \kappa(B \mid A))$ is obviously a negative ranking function in turn and called the *conditionalization of* κ *by* A.

We might rewrite this definition as a law:

$$\kappa(A \cap B) = \kappa(A) + \kappa(B \mid A) \quad [\textit{the law of conjunction (for negative ranks)}]. \tag{17.8}$$

This amounts to the highly intuitive assertion that one has to add the degree of disbelief in B given A to the degree of disbelief in A in order to get the degree of disbelief in A-and-B.

Moreover, it immediately follows for all $A, B \in \mathcal{A}$ with $\kappa(A) < \infty$:

$$\kappa(B \mid A) = 0 \text{ or } \kappa(\overline{B} \mid A) = 0 \quad [\textit{conditional law of negation}]. \tag{17.9}$$

This law says that even conditional belief must be consistent. If both, $\kappa(B \mid A)$ and $\kappa(\overline{B} \mid A)$, were > 0, both, B and $\overline{B}$, would be believed given A, and this ought to be excluded, as long as the condition A itself is considered possible.

Indeed, my favorite axiomatization of ranking theory runs reversely, it consists of the definition of conditional ranks and the conditional law of negation. The latter says that $\min\{\kappa(A \mid A \cup B), \kappa(B \mid A \cup B)\} = 0$, and this and the definition of conditional ranks entail that $\min\{\kappa(A), \kappa(B)\} = \kappa(A \cup B)$, i.e., the law of disjunction. Hence, the only substantial assumption written into ranking functions is conditional consistency, and it is interesting to see that this entails deductive closure as well. Huber (2007) has further improved upon this important idea and shown that ranking theory is indeed nothing but the assumption of dynamic consistency, i.e., the preservation of consistency under any dynamics of belief. (He parallels, in a way, the dynamic Dutch book argument for probabilities by replacing its assumption of no sure loss by the assumption of consistency under all circumstances.)

It is instructive to look at the positive counterpart of negative conditional ranks. If π is the positive ranking function corresponding to the negative ranking function κ, Definition 17.6 simply translates into: $\pi(B \mid A) = \pi(\overline{A} \cup B) - \pi(\overline{A})$. Defining $A \rightarrow B = \overline{A} \cup B$ as set-theoretical 'material implication', we may as well write:

$$\pi(A \rightarrow B) = \pi(B \mid A) + \pi(\overline{A}) \quad [\textit{the law of material implication}]. \tag{17.10}$$

Again, this is highly intuitive. It says that the degree of belief in the material implication $A \rightarrow B$ is added up from the degree of belief in its vacuous truth (i.e., in $\overline{A}$) and the conditional degree of belief of B given A.[9] However, again comparing the negative and the positive version, one can already sense the analogy between probability and ranking theory from (17.8), but hardly from (17.10). This analogy will play a great role in the following subsections.

[9] Thanks again to Matthias Hild for pointing this out to me.

Two-sided ranks have a conditional version as well; it is straightforward. If τ is the two-sided ranking function corresponding to the negative κ and the positive π, then we may simply define:

$$\tau(B \mid A) = \pi(B \mid A) - \kappa(B \mid A) = \kappa(\overline{B} \mid A) - \kappa(B \mid A). \tag{17.11}$$

It will sometimes be useful to refer to these two-sided conditional ranks.

For illustration of negative conditional ranks, let us briefly return to our example of Tweetie. Above, I already mentioned various examples of if-then sentences, some held vacuously true and some non-vacuously. Now we can see that precisely the if-then sentences non-vacuously held true correspond to conditional beliefs. According to the κ specified, you believe, e.g., that Tweetie can fly given it is a bird (since $\kappa(\overline{F} \mid B) = 1$)) and also given it is a bird, but not a penguin (since $\kappa(\overline{F} \mid B\&\overline{P}) = 2$)), that Tweetie cannot fly given it is a penguin (since $\kappa(F \mid P) = 3$) and even given it is a penguin, but not a bird (since $\kappa(F \mid \overline{B}\&P) = 3$)). You also believe that it is not a penguin given it is a bird (since $\kappa(P \mid B) = 1$) and that it is a bird given it is a penguin (since $\kappa(\overline{B} \mid P) = 7$)). And so forth.

Let us now unfold the power of conditional ranks and their relevance to the dynamics of belief in several steps.

Reasons and Their Balance

The first application of conditional ranks is in the theory of confirmation. Basically, Carnap (1950) told us, confirmation is positive relevance. This idea can be explored probabilistically, as Carnap did. But here the idea works just as well. A proposition A confirms or supports or speaks for a proposition B, or, as I prefer to say, A is a reason for B, if A strengthens the belief in B, i.e., if B is more strongly believed given A than given $\overline{A}$, i.e., iff A is positively relevant for B. This is easily translated into ranking terms:

Definition 17.7 Let κ be a negative ranking function for $\mathcal{A}$ and τ the associated two-sided ranking function. Then $A \in \mathcal{A}$ is a *reason for* $B \in \mathcal{A}$ w.r.t. κ iff $\tau(B \mid A) > \tau\big(B \mid \overline{A}\big)$, i.e., iff $\kappa(\overline{B} \mid A) > \kappa(\overline{B} \mid \overline{A})$ or $\kappa(B \mid A) < \kappa(B \mid \overline{A})$.

If P is a standard probability measure on $\mathcal{A}$, then probabilistic positive relevance can be expressed by $P(B \mid A) > P(B)$ or by $P(B \mid A) > P\big(B \mid \overline{A}\big)$. As long as all three terms involved are defined, the two inequalities are equivalent. Usually, then, the first inequality is preferred because its terms may be defined while not all terms of the second inequality are defined. If P is a Popper measure, this argument does not hold, and then it is easily seen that the second inequality is more adequate, just as in the case of ranking functions.[10]

[10] A case in point is the so-called problem of old evidence, which has a simple solution in terms of Popper measures and the second inequality; cf. Joyce (1999, pp. 203ff.).

Confirmation or support may take four different forms relative to ranking functions, which are unfolded in

Definition 17.8 Let κ be a negative ranking function for $\mathcal{A}$, τ the associated two-sided ranking function, and $A, B \in \mathcal{A}$. Then

$$A \text{ is a } \left\{\begin{array}{c} additional \\ sufficient \\ necessary \\ insufficient \end{array}\right\} \text{ reason for } B \text{ w.r.t. } \kappa \text{ iff } \left\{\begin{array}{l} \tau(B \mid A) > \tau\big(B \mid \overline{A}\big) > 0 \\ \tau(B \mid A) > 0 \geq \tau\big(B \mid \overline{A}\big) \\ \tau(B \mid A) \geq 0 > \tau\big(B \mid \overline{A}\big) \\ 0 > \tau(B \mid A) > \tau\big(B \mid \overline{A}\big) \end{array}\right\}.$$

If A is a reason for B, it must obviously take one of these four forms; and the only way to have two forms at once is by being a necessary and sufficient reason.[11]

Talking of reasons here is, I find, natural, but it stirs a nest of vipers. There is a host of philosophical literature pondering about reasons, justifications, etc. Of course, this is a field where multifarious philosophical conceptions clash, and it is not easy to gain an overview over the fighting parties. Here is not the place for starting a philosophical argument[12], but by using the term 'reason' I want at least to submit the claim that the topic may gain enormously by giving a central place to the above explication of reasons.

To elaborate only a little bit: When philosophers feel forced to make precise their notion of a (theoretical, not practical) reason, they usually refer to the notion of a *deductive* reason, as fully investigated in deductive logic. The deductive reason relation is reflexive, transitive, and not symmetric. By contrast, Definition 17.7 captures the notion of a *deductive or inductive* reason. The relation embraces the deductive relation, but it is reflexive, symmetric, and not transitive. Moreover, the fact that reasons may be additional or insufficient reasons according to Definition 17.8 has been neglected by the relevant discussion, which was rather occupied with necessary and/or sufficient reasons. Pursue, though, the use of the latter terms throughout the history of philosophy. Their deductive explication is standard and almost always fits. Often, it is clear that the novel inductive explication given by Definition 17.8 would be inappropriate. Very often, however, the texts are open to that inductive explication as well, and systematically trying to reinterpret these old texts would yield a highly interesting research program in my view.

The topic is obviously inexhaustible. Let me take up only one further aspect. Intuitively, we weigh reasons. This is a most important activity of our mind. We do not only weigh practical reasons in order to find out what to do, we also weigh theoretical reasons. We are wondering whether or not we should believe B, we are searching for reasons speaking in favor or against B, we are weighing these reasons, and we hopefully reach a conclusion. I am certainly not denying the phenomenon of

[11] In earlier publications I spoke of weak instead of insufficient reasons. Thanks to Arthur Merin who suggested the more appropriate term to me.

[12] I attempted to give a partial overview and argument in Spohn (2001a).

inference that is also important, but what is represented as an inference often rather takes the form of such a weighing procedure. 'Reflective equilibrium' is a familiar and somewhat more pompous metaphor for the same thing.

If the balance of reasons is such a central phenomenon the question arises: how can epistemological theories account for it? The question is less well addressed than one should think. However, the fact that there is a perfectly natural Bayesian answer is a very strong and more or less explicit argument in favor of Bayesianism. Let us take a brief look at how that answer goes:

Let P be a (subjective) probability measure over $\mathcal{A}$ and let B be the focal proposition. Let us look at the simplest case, consisting of one reason A for B and the automatic counter-reason $\overline{A}$ against B. Thus, in analogy to Definition 17.7, $P(B \mid A) > P\big(B \mid \overline{A}\big)$. How does P balance these reasons and thus fit in B? The answer is simple, we have:

$$P(B) = P(B \mid A) \cdot P(A) + P\big(B \mid \overline{A}\big) \cdot P\big(\overline{A}\big)\,. \tag{17.12}$$

This means that the probabilistic balance of reason is a *beam balance* in the literal sense. The length of the lever is $P(B \mid A) - P\big(B \mid \overline{A}\big)$; the two ends of the lever are loaded with the *weights* $P(A)$ and $P(\overline{A})$ of the reasons; $P(B)$ divides the lever into two parts of length $P(B \mid A) - P(B)$ and $P(B) - P\big(B \mid \overline{A}\big)$ representing the *strength* of the reasons; and then $P(B)$ must be chosen so that the beam is in balance. Thus interpreted (17.12) is nothing but the law of levers.

Ranking theory has an answer, too, and I am wondering who else has. According to ranking theory, the balance of reasons works like a *spring balance*. Let κ be a negative ranking function for $\mathcal{A}$, τ the corresponding two-sided ranking function, B the focal proposition, and A a reason for B. So, $\tau(B \mid A) > \tau\big(B \mid \overline{A}\big)$. Again, it easily proved that always $\tau(B \mid A) \geq \tau(B) \geq \tau\big(B \mid \overline{A}\big)$. But where in between is $\tau(B)$ located? A little calculation shows the following specification to be correct:

$$\begin{aligned}
&\text{Let } x = \kappa(B \mid \overline{A}) - \kappa(B \mid A) \text{ and } y = \kappa(\overline{B} \mid A) - \kappa(\overline{B} \mid \overline{A}). \textit{ Then}\\
&\text{(a) } x, y \geq 0 \text{ and } \tau(B \mid A) - \tau\big(B \mid \overline{A}\big) = x + y,\\
&\text{(b) } \tau(B) = \tau\big(B \mid \overline{A}\big)\text{, if } \tau(A) \leq -x,\\
&\text{(c) } \tau(B) = \tau(B \mid A)\text{, if } \tau(A) \geq y,\\
&\text{(d) } \tau(B) = \tau(A) + \tau\big(B \mid \overline{A}\big) + x, \text{ if } -x < \tau(A) < y.
\end{aligned} \tag{17.13}$$

This does not look as straightforward as the probabilistic beam balance. Still, it is not so complicated to interpret (17.13) as a spring balance. The idea is that you hook in the spring at a certain point, that you extend it by the force of reasons, and that $\tau(B)$ is where the spring extends. Consider first the case where $x, y > 0$. Then you hook in the spring at point 0 $\big(= \tau\big(B \mid \overline{A}\big) + x\big)$ and exert the force $\tau(A)$ on the spring. Either, this force transcends the lower stopping point $-x$ or the upper stopping point y. Then the spring extends exactly till the stopping point, as (17.13b + c) say. Or, the force $\tau(A)$ is less. Then the spring extends exactly by $\tau(A)$, according to (17.13d).

The second case is that $x = 0$ and $y > 0$. Then you fix the spring at $\tau\left(B \mid \overline{A}\right)$, the lower point of the interval in which $\tau(B)$ can move. The spring cannot extend below that point, says (17.13b). But according to (17.13c + d) it can extend above, by the force $\tau(A)$, but not beyond the upper stopping point. For the third case $x > 0$ and $y = 0$ just reverse the second picture. In this way, the force of the reason A, represented by its two-sided rank $\tau(A)$, pulls the two-sided rank of the focal proposition B to its proper place within the interval $\left[\tau\left(B \mid \overline{A}\right), \tau(B \mid A)\right]$ fixed by the relevant conditional ranks.

I do not want to assess these findings in detail. You might prefer the probabilistic balance of reasons, a preference I would understand. You might be happy to have at least one alternative model, an attitude I recommend. Or you may search for further models of the weighing of reasons; in this case, I wish you good luck. What you may not do is ignoring the issue; your epistemology is incomplete if it does not take a stance. And one must be clear about what is required for taking a stance. As long as one considers positive relevance to be the basic characteristic of reasons, one must provide some notion of conditional degrees of belief, conditional probabilities, conditional ranks, or whatever. Without some well-behaved conditionalization one cannot succeed.

The Dynamics of Belief and the Measurement of Belief

Our next point will be to define a reasonable dynamics for ranking functions that entails a dynamic for belief. There are many causes which affect our beliefs, forgetfulness as a necessary evil, drugs as an unnecessary evil, and so on. From a rational point of view, it is scarcely possible to say anything about such changes.[13] The rational changes are due to experience or information. Thus, it seems we have already solved our task: if κ is my present doxastic state and I get informed about the proposition A, then I move to the conditionalization κ_A of κ by A. This, however, would be a bad idea. Recall that we have $\kappa_A\left(\overline{A}\right) = \infty$, i.e., A is believed with absolute certainty in κ_A; no future evidence could cast any doubt on the information. This may sometimes happen; but usually information does not come so firmly. Information may turn out wrong, evidence may be misleading, perception may be misinterpreted; we should provide for flexibility. How?

One point of our first attempt was correct; if my information consists solely in the proposition A, this cannot affect my beliefs conditional on A. Likewise it cannot affect my beliefs conditional on $\overline{A}$. Thus, it directly affects only how firmly I believe A itself. So, how firmly should I believe A? There is no general answer. I propose to turn this into a parameter of the information process itself; somehow the way I get informed about A entrenches A in my belief state with a certain firmness x.

[13] Although there is a (by far not trivial) decision rule telling that costless memory is never bad, just as costless information; cf. Spohn (1976/78, sect. 4.4).

The point is that as soon as the parameter is fixed and the constancy of the relevant conditional beliefs is accepted, my posterior belief state is fully determined. This is the content of

Definition 17.9 Let κ be a negative ranking function for $\mathcal{A}$, $A \in \mathcal{A}$ such that $\kappa(A)$, $\kappa(\overline{A}) < \infty$, and $x \in \mathbf{R}^*$. Then the $A \to x$-conditionalization $\kappa_{A\to x}$ of κ is defined by $\kappa_{A\to x}(B) = \begin{cases} \kappa(B \mid A) \text{ for } B \subseteq A, \\ \kappa(B \mid \overline{A}) + x \text{ for } B \subseteq \overline{A} \end{cases}$. From this $\kappa_{A\to x}(B)$ may be inferred for all other $B \in \mathcal{A}$ with the law of disjunction.

Hence, the effect of the $A \to x$-conditionalization is to shift the possibilities in A (to lower ranks) so that $\kappa_{A\to x}(A) = 0$ and the possibilities in $\overline{A}$ (to higher ranks) so that $\kappa_{A\to x}(\overline{A}) = x$. If one is attached to the idea that evidence consists in nothing but a proposition, the additional parameter is a mystery. The processing of evidence may indeed be so automatic that one hardly becomes aware of this parameter. Still, I find it entirely natural that evidence comes more or less firmly. Consider, for instance, the proposition: "There are tigers in the Amazon jungle", and consider six scenarios: (a) I read a somewhat sensationalist coverage in the yellow press claiming this, (b) I read a serious article in a serious newspaper claiming this, (c) I hear the Brazilian government officially announcing that tigers have been discovered in the Amazon area, (d) I see a documentary in TV claiming to show tigers in the Amazon jungle, (e) I read an article in *Nature* by a famous zoologist reporting of tigers there, (f) I travel there by myself and see the tigers. In all six cases I receive the information that there are tigers in the Amazon jungle, but with varying and, I find, increasing certainty.

One might object that the evidence and thus the proposition received is clearly a different one in each of the scenarios. The crucial point, though, is that we are dealing here with a fixed algebra $\mathcal{A}$ of propositions and that we have nowhere presupposed that this algebra consists of all propositions whatsoever; indeed, that would be a doubtful presupposition. Hence $\mathcal{A}$ may be course-grained and unable to represent the propositional differences between the scenarios; the proposition in $\mathcal{A}$ which is directly affected in the various scenarios may be just the proposition that there are tigers in the Amazon jungle. Still the scenarios may be distinguished by the firmness parameter.

So, the dynamics of ranking functions I propose is simply this: Suppose κ is your prior doxastic state. Now you receive some information A with firmness x. Then your posterior state is $\kappa_{A\to x}$. Your beliefs change accordingly; they are what they are according to $\kappa_{A\to x}$. Note that the procedure is iterable. Next, you receive the information B with firmness y, and so you move to $(\kappa_{A\to x})_{B\to y}$. And so on. This point will acquire great importance later on.

I should mention, though, that this iterability need not work in full generality. Let us call a negative ranking function κ *regular* iff $\kappa(A) < \infty$ for all $A \neq \varnothing$. Then we obviously have that $\kappa_{A\to x}$ is regular if κ is regular and $x < \infty$. Within the realm of regular ranking functions iteration of changes works without restriction. Outside this realm you may get problems with the rank ∞.

There is an important generalization of Definition 17.9. I just made a point of the fact that the algebra $\mathcal{A}$ may be too coarse-grained to propositionally represent all possible evidence. Why assume then that it is just one proposition A in the algebra that is directly affected by the evidence? Well, we need not assume this. We may more generally assume that the evidence affects some evidential partition $\mathcal{E}$ of W and assigns some new ranks to the members of the partition, which we may sum up in a complete ranking function λ on $\mathcal{E}$. Then we may define the $\mathcal{E} \to \lambda$-*conditionalization* $\kappa_{\mathcal{E}\to\lambda}$ of the prior κ by $\kappa_{\mathcal{E}\to\lambda}(B) = \kappa(B \mid E_i) + \lambda(E_i)$ for $B \subseteq E_i$ $(i = 1, \ldots, n)$ and infer $\kappa_{\mathcal{E}\to\lambda}(B)$ for all other B by the law of disjunction. This is the most general law of doxastic change in terms of ranking functions I can conceive of. Note that we may describe the $\mathcal{E} \to \lambda$-conditionalization of κ as the mixture of all κ_{E_i} $(i = 1, \ldots, n)$. So, this is a first useful application of mixtures of ranking functions.

Here, at last, the reader will have noticed the great similarity of my conditionalization rules with Jeffrey's probabilistic conditionalization first presented in Jeffrey (1965, ch. 11). Indeed, I have completely borrowed my rules from Jeffrey. Still, let us further defer the comparison of ranking with probability theory. The fact that many things run similarly does not mean that one can dispense with the one in favor of the other, as I shall make clear in section "Ranks and probabilities", pp. 328ff.

There is an important variant of Definition 17.9. Shenoy (1991), and several authors after him, pointed out that the parameter x as conceived in Definition 17.9 does not characterize the evidence as such, but rather the result of the interaction between the prior doxastic state and the evidence. Shenoy proposed a reformulation with a parameter exclusively pertaining to the evidence:

Definition 17.10 Let κ be a negative ranking function for $\mathcal{A}$, $A \in \mathcal{A}$ such that $\kappa(A)$, $\kappa(\overline{A}) < \infty$, and $x \in \mathbf{R}^*$. Then the $A{\uparrow}x$-*conditionalization* $\kappa_{A\uparrow x}$ of κ is defined by $\kappa_{A\uparrow x}(B) = \begin{cases} \kappa(B) - y \text{ for } B \subseteq A, \\ \kappa(B) + x - y \text{ for } B \subseteq \overline{A}, \end{cases}$ where $y = \min\{\kappa(A), x\}$. Again, $\kappa_{A\uparrow x}(B)$ may be inferred for all other $B \in \mathcal{A}$ by the law of disjunction.

The effect of this conditionalization is easily stated. It is, whatever the prior ranks of A and $\overline{A}$ are, that the possibilities within A improve by exactly x ranks in comparison to the possibilities within $\overline{A}$. In other words, we always have $\tau_{A\uparrow x}(A) - \tau(A) = x$ (in terms of the prior and the posterior two-sided ranking function).

It is thus appropriate to say that in $A{\uparrow}x$-conditionalization the parameter x exclusively characterizes the evidential impact. We may characterize the $A{\to}x$-conditionalization of Definition 17.9 as *result-oriented* and the $A{\uparrow}x$-conditionalization of Definition 17.10 as *evidence-oriented*. Of course, the two variants are easily interdefinable. We always have $\kappa_{A\to x} = \kappa_{A\uparrow y}$, where $y = x - \tau(A) = x + \tau(\overline{A})$. Still, it is sometimes useful to change perspective from one variant to the other.[14]

[14] Generalized probabilistic conditionalization as originally proposed by Jeffrey was result-oriented as well. However, Garber (1980) observed that there is also an evidence-oriented version of generalized probabilistic conditionalization. The relation, though, is not quite as elegant.

For instance, the evidence-oriented version helps to some nice observations. We may note that conditionalization is reversible: $(\kappa_{A\uparrow x})_{\overline{A}\uparrow x} = \kappa$. So, there is always a possible second change undoing the first. Moreover, changes always commute: $(\kappa_{A\uparrow x})_{B\uparrow_y} = (\kappa_{B\uparrow y})_{A\uparrow x}$. In terms of result-oriented conditionalization this law would look more awkward. Commutativity does not mean, however, that one could comprise the two changes into a single change. Rather, the joint effect of two conditionalizations according to Definition 17.9 or 17.10 can in general only be summarized as one step of generalized $\mathcal{E} \rightarrow \lambda$-conditionalization. I think that reversibility and commutativity are intuitively desirable.

Change through conditionalization is driven by information, evidence, or perception. This is how I have explained it. However, we may also draw a more philosophical picture, we may also say that belief change according to Definition 17.9 or 17.10 is driven by reasons. Propositions for which the information received is irrelevant do not change their ranks, but propositions for which that information is positively or negatively relevant do change their ranks. The evidential force pulls at the springs and they must find a new rest position for all the propositions for or against which the evidence speaks, just in the way I have described in the previous subsection.

This is a strong picture captivating many philosophers. However, I have implemented it in a slightly unusual way. The usual way would have been to attempt to give some substantial account of what reasons are on which an account of belief dynamics is thereafter based. I have reversed the order. I have first defined conditionalization in Definition 17.6 and the more sophisticated form in Definitions 17.9 and 17.10. With the help of conditionalization, i.e., from this account of belief dynamics, I could define the reason relation in a way sustaining this picture. At the same time this procedure entails dispensing with a more objective notion of a reason. Rather, what is a reason for what is entirely determined by the subjective doxastic state as represented by the ranking function at hand. Ultimately, this move is urged by inductive skepticism as enforced by David Hume and reinforced by Nelson Goodman. But it does not mean surrender to skepticism. On the contrary, we are about to unfold a positive theory of rational belief and rational belief change, and we shall have to see how far it carries us.[15]

If one looks at the huge literature on belief change, one finds discussed predominantly three kinds of changes: expansions, revisions, and contractions. Opinions widely diverge concerning these three kinds. For Levi, for instance, revisions are whatever results form concatenating contractions and expansions according to the so-called Levi identity, and so he investigates the latter (see his most recent account in Levi 2004). The AGM approach characterizes both, revisions and contractions, and claims nice correspondences back and forth by help of the Levi and the Harper identity (cf., e.g., Gärdenfors 1988, chs. 3 and 4). Or one might object to the characterization of contraction, but accept that of revision, and hence reject these identities. And so forth.

[15]Here it does not carry us far beyond the beginnings. In Spohn (1991, 1999) I have argued for some stronger rationality requirements and their consequences.

I do not really want to discuss the issue. I only want to point out that we have already taken a stance insofar as expansions, revisions, and contractions are all special cases of our $A \to x$–conditionalization. This is more easily explained in terms of result-oriented conditionalization:

If $\kappa(A) = 0$, i.e., if A is not disbelieved, then $\kappa_{A \to x}$ represents an *expansion* by A for any $x > 0$. If $\kappa(\overline{A}) = 0$, the expansion is genuine, if $\kappa(\overline{A}) > 0$, i.e., if A is already believed in κ, the expansion is vacuous. Are there many different expansions? Yes and no. Of course, for each $x > 0$ a different $\kappa_{A \to x}$ results. On the other hand, one and the same belief set is associated with all these expansions. Hence, the expanded belief set is uniquely determined.

Similarly for revision. If $\kappa(A) > 0$, i.e., if A is disbelieved, then $\kappa_{A \to x}$ represents a genuine *revision* by A for any $x > 0$. In this case, the belief in $\overline{A}$ must be given up and along with it many other beliefs; instead, A must be adopted together with many other beliefs. Again, there are many different revisions, but all of them result in the same revised belief set.

Finally, if $\kappa(A) = 0$, i.e., if A is not disbelieved, then $\kappa_{A \to 0}$ represents contraction by A. If $\kappa(\overline{A}) > 0$, i.e., if A is even believed, the contraction is genuine; then belief in A is given up after contraction and no new belief adopted. If $\kappa(\overline{A}) = 0$, the contraction is vacuous; there was nothing to contract in the first place. If $\kappa(A) > 0$, i.e., if $\overline{A}$ is believed, then $\kappa_{A \to 0} = \kappa^{-}_{\overline{A} \to 0}$ rather represents contraction by $\overline{A}$.[16]

As observed in Spohn (1988, footnote 20) and more fully explained in Gärdenfors (1988, pp. 73f.), it is easily checked that expansions, revisions, and contractions thus defined satisfy all of the original AGM postulates (K*1-8) and (K^{-}1-8) (cf. Gärdenfors 1988, pp. 54–56 and 61–64) (when they are translated from AGM's sentential framework into our propositional or set-theoretical one). For those like me who accept the AGM postulates this is a welcome result.

For the moment, though, it may seem that we have simply reformulated AGM belief revision theory. This is not so; $A \to x$-conditionalization is much more general than the three AGM changes. This is clear from the fact that there are many different expansions and revisions that cannot be distinguished by the AGM account. It is perhaps clearest in the case of vacuous expansion that is no change at all in the AGM framework, but may well be a genuine change in the ranking framework, a redistribution of ranks which does not affect the surface of beliefs. Another way to state the same point is that insufficient and additional reasons also drive doxastic changes, which, however, are inexpressible in the AGM framework. For instance, if A is still disbelieved in the *$A \uparrow x$-conditionalization* $\kappa_{A \uparrow x}$ of κ (since $\kappa(A) > x$), one has obviously received only an insufficient reason for A, and the *$A \uparrow x$-conditionalization* might thus be taken to represent what is called non-prioritized belief revision in the AGM literature (cf. Hansson 1997).

[16] If we accept the idea in section "Basics" (p. 311) of taking the interval $[-z, z]$ of two-sided ranks as the range of neutrality, contraction seems to become ambiguous as well. However, the contraction just defined would still be distinguishable as a *central* contraction since it gives the contracted proposition central neutrality.

This is not the core of the matter, though. The core of the matter is *iterated belief change*, which I have put into the center of my considerations in Spohn (1983, sect. 5.3, 1988). As I have argued there, AGM belief revision theory is essentially unable to account for iterated belief change. I take 20 years of multifarious, but in my view unsatisfactory attempts to deal with that problem (see the overview in Rott 2008) as confirming my early assessment. By contrast, changes of the type $A \to x$-conditionalization are obviously indefinitely iterable.

In fact, my argument in Spohn (1988) was stronger. It was that if AGM belief revision theory is to be improved so as to adequately deal with the problem of iterated belief change, ranking theory is the only way to do it. I always considered this to be a conclusive argument in favor of ranking theory.

This may be so. Still, AGM theorists, and others as well, remained skeptical. "What exactly is the meaning of numerical ranks?" they asked. One may well acknowledge that the ranking apparatus works in a smooth and elegant way, has a lot of explanatory power, etc. But all this does not answer this question. Bayesians have met this challenge. They have told stories about the operational meaning of subjective probabilities in terms of betting behavior, they have proposed an ingenious variety of procedures for measuring this kind of degrees of belief. One would like to see a comparative achievement for ranking theory.

It exists and is finally presented in Hild and Spohn (2008). There is no space here to fully develop the argument. However, the basic point can easily be indicated so as to make the full argument at least plausible. The point is that ranks do not only account for iterated belief change, but can reversely be measured thereby. This may at first sound unhelpful. $A \to x$-conditionalization refers to the number x; so even if ranks can somehow be measured with the help of such conditionalizations, we do not seem to provide a fundamental measurement of ranks. Recall, however, that (central) contraction by A (or $\overline{A}$) is just $A \to 0$-conditionalization and is thus free of a hidden reference to numerical ranks; it only refers to rank 0 which has a clear operational or surface interpretation in terms of belief. Hence, the idea is to measure ranks by means of iterated contractions; if that works, it really provides a fundamental measurement of ranks that is based only on the beliefs one now has and one would have after various iterated contractions.

How does the idea work? Recall our observation above that the positive rank of a material implication $A \to B$ is the sum of the degree of belief in B given A and the degree of belief in the vacuous truth of $A \to B$, i.e., of $\overline{A}$. Hence, after contraction by $\overline{A}$, belief in the material implication $A \to B$ is equivalent to belief in B given A, i.e., to the positive relevance of A to B. This is how the reason relation, i.e., positive relevance, manifests itself in beliefs surviving contractions. Similarly for negative relevance and irrelevance.

Next observe that positive relevance can be expressed by certain inequalities for ranks that compare certain differences between ranks (similarly for negative relevance and irrelevance). This calls for applying the theory of difference measurement, as paradigmatically presented by Krantz et al. (1971, ch. 4).

Let us illustrate how this might work in our Tweetie example, pp. 306f. There we had specified a ranking function κ for the eight propositional atoms, entailing ranks

for all 256 propositions involved. Focusing on the atoms, we are thus dealing with a realm $X = \{x_1, \ldots x_8\}$ (where $x_1 = B \,\&\, \overline{P} \,\&\, F$, etc.) and a numerical function f such that

$$f(x_1) = 0,\ f(x_2) = 4,\ f(x_3) = 0,\ f(x_4) = 11,$$
$$f(x_5) = 2,\ f(x_6) = 1,\ f(x_7) = 0,\ f(x_8) = 8.$$

This induces a lot of difference comparisons. For instance, we have, $f(x_6) - f(x_5) < f(x_2) - f(x_1)$. It is easily checked that this inequality says that, given B (being a bird), P (being a penguin) is positively relevant to $\overline{F}$ (not being able to fly) and that this in turn is equivalent with $P \rightarrow \overline{F}$ or $\overline{P} \rightarrow F$ still being believed after iterated contraction first by $\overline{B}$ and then by P and $\overline{P}$ (only one of the latter is a genuine contraction). Or we have $f(x_2) - f(x_6) = f(x_4) - f(x_8)$. Now, this is an equality saying that, given P, B (and $\overline{B}$) is irrelevant to F (and $\overline{F}$), and this in turn is equivalent with none of the four material implications from B or $\overline{B}$ to F or $\overline{F}$ being believed after iterated contraction first by $\overline{P}$ and then by B and $\overline{B}$ (again, only one of the latter is a genuine contraction).

Do these comparisons help to determine f? Yes, the example was so constructed: First, we have $f(x_1) - f(x_3) = f(x_3) - f(x_1) = f(x_1) - f(x_7)$. This entails $f(x_1) = f(x_3) = f(x_7)$. Let us choose this as the zero point of our ranking scale; i.e., $f(x_1) = 0$. Next, we have $f(x_5) - f(x_6) = f(x_6) - f(x_1)$. If we choose $f(x_6) = 1$ as our ranking unit, this entails $f(x_5) = 2$. Then, we have $f(x_2) - f(x_5) = f(x_5) - f(x_1)$, entailing $f(x_2) = 4$, and $f(x_8) - f(x_2) = f(x_2) - f(x_1)$, entailing $f(x_8) = 8$. Finally, we have $f(x_4) - f(x_8) = f(x_2) - f(x_6)$, the equation I had already explained, so that $f(x_4) = 11$. In this way, the difference comparisons entailed by our specification of f determine f uniquely up to a unit and a zero point.

The theory of difference measurement tells us how this procedure works in full generality. The resulting theorem says the following: Iterated contractions behave thus and thus if and only if differences between ranks behave thus and thus; and if differences between ranks behave thus and thus, then there is a ranking function measured on a ratio scale, i.e., unique up to a multiplicative constant, which exactly represents these differences. (See theorems 4.12 and 6.21 in Hild and Spohn (2008) for what "thus and thus" precisely means.)

On the one hand, this provides for an axiomatization of iterated contraction going beyond Darwiche and Pearl (1997), who presented generally accepted postulates of iterated revision and contraction and partially agreeing and partially disagreeing with further postulates proposed.[17] This axiomatization is assessible on intuitive and other grounds. On the other hand, one knows that if one accepts this axiomatization of iterated contraction one is bound to accept ranks as I have proposed them. Ranks do not fall from the sky, then; on the contrary, they uniquely represent contraction behavior.

[17] For an overview over such proposals see Rott (2008). For somewhat more detailed comparative remarks see Hild and Spohn (2008, sect. 5).

Conditional Independence and Bayesian Nets

It is worthwhile looking a bit more at the details of belief formation and revision. For this purpose we should give more structure to propositions. They have a Boolean structure so far, but we cannot yet compose them from basic propositions as we intuitively do. A common formal way to do this is to generate propositions from (random) variables. I identify a variable with the set of its possible values. I intend variables to be specific ones. E.g., the temperature at March 15, 2005, in Konstanz (not understood as the actual temperature, but as whatever it may be, say, between -100 and $+100$ °C) is such a variable. Or, to elaborate, if we consider each of the six general variables temperature, air pressure, wind, humidity, precipitation, cloudiness at each of the 500 weather stations in Germany twice a day at each of the 366 days of 2004, we get a collection of $6 \times 500 \times 732$ specific variables with which we can draw a detailed picture of the weather in Germany in 2004.

So, let V be the set of specific variables considered, where each $v \in V$ is just at least a binary set. A possible course of events or a possibility, for short, is just a selection function w for V, i.e., a function w on V such that $w(v) \in v$ for all $v \in V$. Hence, each such function specifies a way how the variables in V may realize. The set of all possibilities then simply is $W = \bigtimes V$. As before, propositions are subsets of W. Now, however, we can say that propositions are *about* certain variables. Let $X \subseteq V$. Then we say that $w, w' \in W$ *agree on* X iff $w(v) = w'(v)$ for all $v \in X$. And we define that a proposition A is *about* $X \subseteq V$ iff, for each w in A, all w' agreeing with w on X are in A as well. Let $\mathcal{A}(X)$ be the set of propositions about X. Clearly, $\mathcal{A}(X) \subseteq \mathcal{A}(Y)$ for $X \subseteq Y$, and $\mathcal{A} = \mathcal{A}(V)$. In this way, propositions are endowed with more structure. We may conceive of propositions about single variables as *basic* propositions; the whole algebra $\mathcal{A}$ is obviously generated by such basic propositions (at least if V is finite). So much as preparation for the next substantial step.

This step consists in more closely attending to (doxastic) dependence and independence in ranking terms. In a way, we have already addressed this issue: dependence is just positive or negative relevance, and independence is irrelevance. Still, let me state

Definition 17.11 Let κ be a negative ranking function for $\mathcal{A}$ and $A, B, C \in \mathcal{A}$. Then A and B are *independent* w.r.t. κ, i.e., $A \perp B$, iff $\tau(B \mid A) = \tau\big(B \mid \overline{A}\big)$, i.e., iff for all $A' \in \big\{A, \overline{A}\big\}$ and $B' \in \big\{B, \overline{B}\big\}$ $\kappa(A' \cap B') = \kappa(A') + \kappa(B')$. And A and B are *independent given* C w.r.t. κ, i.e., $A \perp B \,/\, C$, iff A and B are independent w.r.t. κ_C.

(Conditional) independence is symmetric. If A is independent from B, $\overline{A}$ is so as well. If A is independent from B and A' disjoint from A, then A' is independent from B iff $A \cup A'$ is. $\varnothing$ and W are independent from all propositions. And so on.

The more interesting notion, however, is dependence and independence among variables. Look at probability theory where research traditionally and overwhelmingly focused on independent series of random variables and on Markov processes that are characterized by the assumption that past and future variables are

independent given the present variable. We have already prepared for explaining this notion in ranking terms as well.

Definition 17.12 Let κ be a ranking function for $\mathcal{A} = \mathcal{A}(V)$, and let $X, Y, Z \subseteq V$ be sets of variables. Then X and Y are *independent* w.r.t. κ, i.e., $X \perp Y$, iff $A \perp B$ for all $A \in \mathcal{A}(X)$ and all $B \in \mathcal{A}(Y)$. Let moreover $\mathcal{Z}(Z)$ be the set of atoms of $\mathcal{A}(Z)$, i.e., the set of the logically strongest, non-empty proposition in $\mathcal{A}(Z)$. Then X and Y are *independent given* Z w.r.t. κ, i.e., $X \perp Y / Z$, iff $A \perp B/C$ for all $A \in \mathcal{A}(X)$, $B \in \mathcal{A}(Y)$, and $C \in \mathcal{Z}(Z)$.

In other words, $X \perp Y/Z$ iff all propositions about X are independent from all propositions about Y given any full specification of the variables in Z. Conditional independence among sets of variables obey the following laws:

Let κ be a negative ranking function for $\mathcal{A}(V)$. Then for any mutually disjoint $X, Y, Z, U \subseteq V$:

(a) if $X \perp Y/Z$, then $Y \perp X/Z$ [Symmetry],
(b) if $X \perp Y \cup U/Z$, then $X \perp Y/Z$ and $X \perp U/Z$ [Decomposition],
(c) $X \perp Y \cup U/Z$, then $X \perp Y/Z \cup U$ [Weak Union],
(d) $X \perp Y/Z$ and $X \perp U/Z \cup Y$, then $X \perp Y \cup U/Z$ [Contraction],
(e) if κ is regular and if $X \perp Y/Z \cup U$ and $X \perp U/Z \cup Y$, then $X \perp Y \cup U/Z$ [Intersection]

(17.14)

These are nothing but what Pearl (1988, p. 88) calls the *graphoid* axioms; the labels are his (cf. p. 84). (Note that law (d), contraction, has nothing to do with contraction in belief revision theory.) That probabilistic conditional independence satisfies these laws was first proved in Spohn (1976/78, sect. 3.2) and Dawid (1979). The ranking Theorem (17.14) was proved in Spohn (1983, sect. 5.3, 1988, sect. 6). I conjectured in 1976, and Pearl conjectured, too, that the graphoid axioms give a complete characterization of conditional independence. We were disproved, however, by Studeny (1989) w.r.t. probability measures, but the proof carries over to ranking functions (cf. Spohn 1994a). Under special conditions, though, the graphoid axioms *are* complete, as was proved by Geiger and Pearl (1990) for probability measures and by Hunter (1991) for ranking functions (cf. again, Spohn 1994a).

I am emphasizing all this, because the main purport of Pearl's path-breaking book (1988) is to develop what he calls the theory of Bayesian nets, a theory that has acquired great importance and is presented in many text books (see, e.g., Neapolitan 1990 or Jensen 2001). Pearl makes very clear that the basis of this theory consists in the graphoid axioms; these allow representing conditional dependence and independence among sets of variables by Bayesian nets, i.e., by directed acyclic graphs, the nodes of which are variables. A vertex $u \rightarrow v$ of the graph then represents the fact that v is dependent on u given all the variables preceding v in some given order, for instance, temporally preceding v. A major point of this theory is that it can describe in detail how probabilistic change triggered at some node in the net

propagates throughout the net. All this is not merely mathematics, it is intuitively sensible and philosophically highly significant; for instance, inference acquires a novel and fruitful meaning in the theory of Bayesian nets.

Of course, my point now is that all these virtues carry over to ranking theory with the help of observation (17.14). The point is obvious, but hardly elaborated; that should be done.[18] It will thus turn out that ranks and hence beliefs can also be represented and computationally managed in that kind of structure.

This is not yet the end of the story. Spirtes et al. (1993) (see also Pearl 2000) have made amply clear that probabilistic Bayesian nets have a most natural causal interpretation; a vertex $u \rightarrow v$ then represents that the variable v directly causally depends on the variable u. Spirtes et al. back up this interpretation, i.e., this connection of probability and causality, by their three basic axioms: the causal Markov condition, the minimality condition, and, less importantly, the faithfulness condition (cf. Spirtes et al. 1993, sect. 3.4). And they go on to develop a really impressive account of causation and causal inference on the basis of these axioms and thus upon the theory of Bayesian nets.

Again, all this carries over to ranking theory. Indeed, this is what ranks were designed for in the first place. In Spohn (1983) I gave an explication of probabilistic causation that entails the causal Markov condition and the minimality condition, and also Reichenbach's principle of the common cause, as I observed later in Spohn (1994b).[19] And I was convinced of the idea that, if the theory of causation is bound to bifurcate into a deterministic and a probabilistic branch, these two branches must at least be developed in perfect parallel. Hence, I proposed ranking theory in Spohn (1983) in order to realize this idea.[20] Of course, one has to discuss how adequate that theory of deterministic causation is, just as the adequacy of the causal interpretation of Bayesian nets is open to discussion. Here, my point is only that this deep philosophical perspective lies within reach of ranking theory; it is what originally drove that theory.

Objective Ranks?

Now, a fundamental problem of ranking theory is coming into sight. I have emphasized that ranking functions represent rational beliefs and their rational dynamics and are thus entirely subject-bound. You have your ranking function and I have mine. We may or may not harmonize. In any case, they remain our subjective property.

[18] It has been done in the meantime. See Hohenadel (2013).

[19] I have analyzed the relation between Spirtes' et al. axiomatic approach to causation and my definitional approach a bit more thoroughly in Spohn (2001b).

[20] For a recent presentation of the account of deterministic causation in terms of ranking functions and its comparison in particular with David Lewis' counterfactual approach see Spohn (2006).

I have also emphasized the analogy to probability theory. There, however, we find subjective *and* objective probabilities. There are radicals who deny the one or the other kind of probability; and the nature of objective probabilities may still be ill understood. So, we certainly enter mined area here. Still, the predominant opinion is that both, the subjective and the objective notion, are somehow meaningful.

We therefore face a tension. It increases with our remarks about causation. I said I have provided an analysis of causation in ranking terms. If this analysis were to go through, the consequence would be that causal relations obtain relative to a ranking function, i.e., relative to the doxastic state of a subject. David Hume endorsed and denied this consequence at the same time; he was peculiarly ambiguous. This ambiguity must, however, be seen as his great achievement with which all philosophers after him had and still have to struggle. In any case, it will not do to turn causation simply into a subjective notion, as I seem to propose. If my strategy is to work at all, then the actually existing causal relations have to be those obtaining relative to the objectively correct ranking function. Is there any way to make sense of this phrase? (It is not even a notion yet.)

Yes, partially. The beginning is easy. Propositions are objectively true or false, and so are beliefs. Hence, a ranking function may be called objectively true or false as well, according to the beliefs it embodies. However, this is a very small step. Ranking functions can agree in their belief sets or in the propositions receiving rank 0, and yet widely diverge in the other ranks and thus in inductive and dynamic behavior. So, the suggested beginning is a very small step, indeed.

Taking a bigger step is more difficult. In my (1993) I have made a precise and detailed proposal that I still take to be sound; there is no space to repeat it here. Let me only briefly explain the basic idea. It is simply this: If propositions and beliefs are objectively true or false, then other features of ranking functions can be objectified to the extent to which these features are uniquely reflected in the associated belief sets. One constructive task is then to precisely define the content of the phrase 'uniquely reflected' and the required presuppositions or restrictions. The other constructive task is to inquire which specific features can in this sense be objectified to which specific extent.

Very roughly, the results in my (1993) are this: First, positive relevance, i.e., the reason relation, is *not* objectifiable in this sense, even if restricted to necessary and/or sufficient reasons. Second, whenever A is a sufficient or necessary direct cause of B w.r.t. κ, there is an associated material implication of the form "if the relevant circumstances obtain, then if A, then B, or, respectively, if $\overline{A}$, then $\overline{B}$". I call the conjunction of all these material implications the *causal law* associated with κ. The causal law is a proposition, an objective truth-condition. The point now is that there is a rich class of ranking functions which, under certain presuppositions, can uniquely be reconstructed from their causal laws and which may thus be called causal laws as well. In this sense and to this extent, causal relations obtaining relative to a subjective raking function can be objectified and thus do hold objectively.

A special case treated in Spohn (2002, 2005a) is the case of strict or deterministic laws. A strict law is, by all means, a regularity, an invariable obtaining of a certain type of state of affairs. But not any regularity is a law. What I have proposed in

Spohn (2002) is that a law is an independent and identically distributed (infinite) repetition of the type of state in question or, rather, in order for that phrase to make sense, an independent and identically distributed repetition of a certain ranking assessment of that type of state. Hence, a law is a certain kind of ranking function. This sounds weird, because a law thus turns into a kind of doxastic attitude. The literature on lawlikeness shows, however, that this is not so absurd a direction; if, besides explanatory power or support of counterfactuals, projectibility or inductive behavior are made defining features of laws, they are characterized by their epistemic role and thus get somehow entangled with our subjective states (see also Lange 2000, ch. 7, on the root commitment associated with laws). The main point, though, is that the ranking functions expressing deterministic laws are again of the objectifiable kind. So, there is a way of maintaining even within this account that laws obtain mind-independently.

In fact, according to what I have sketched, a deterministic law is the precise ranking analogue of a statistical law. De Finetti (1937) has proposed an ingenious way of eliminating objective probabilities and statistical laws by showing, in his famous representation theorem, that beliefs (i.e., subjective probabilities) about statistical laws (describing an infinite sequence of independent and identically distributed trials) are strictly equivalent to symmetric or exchangeable subjective probabilities for these trials and that experience makes these symmetric probabilities converge to the true statistical law. The eliminativist intention of the story is mostly dismissed today; rather, objective probabilities are taken seriously. Still, de Finetti's account has remained a paradigm story about the relation between subjective and objective probabilities.

I am mentioning all this because this paradigm story can be directly transferred to ranking theory. Let κ be any ranking function for an infinite sequence of trials (= variables) which is regular and symmetric and according to which the outcome of a certain trial is not negatively relevant to the same outcome in the next trial. Then κ is a unique mixture of deterministic laws for that sequence of trials in the above-mentioned sense, and experience makes κ converge to the true deterministic law. (Cf. Spohn 2005a for all this, where I have treated only the simplest case of the infinite repetition of a binary variable or a trial having only two possible outcomes. With an additional condition, however, the results generalize to all variables taking finitely many values).

This may suffice as an overview over the basics of ranking theory and its elaboration into various directions; it got long enough. In a way, my overall argument in section "Further comparisons", pp. 335ff, when I shall make a bit more detailed comparative remarks about other members of the Baconian probability family, should be clear by now: Bayesian epistemology has enormous powers and virtues and rich details and ramifications. Small wonder that Pascal by far outstripped Bacon. In a nutshell, I have explained that many essential virtues can be duplicated in ranking theory; indeed, the duplications can stand on their own, having an independent significance. Bacon can catch up with Pascal. Of course, my rhetorical question will then be: Which other version of Baconian probability is able to come up with similar results?

Still, one might suspect that I can claim these successes only by turning Bacon into a fake Pascal. I have never left the Bayesian home, it may seem. Hence, one might even suspect that ranking theory is superfluous and may be reduced to the traditional Bayesian point of view. In other words, it is high time to study more closely the relation between probability and ranking theory. This will be our task in the next section.

Ranks and Probabilities

The relation between probabilities and ranks is surprisingly complex and fascinating. I first turn to the more formal aspects of the comparison before discussing the philosophical aspects.

Formal Aspects

The reader will have observed since long why ranks behave so much like probabilities. There is obviously a simple translation of probability into ranking theory: translate the sum of probabilities into the minimum of ranks, the product of probabilities into the sum of ranks, and the quotient of probabilities into the difference of ranks. Thereby, the probabilistic law of additivity turns into the law of disjunction, the probabilistic law of multiplication into the law of conjunction (for negative ranks), and the definition of conditional probabilities into the definition of conditional ranks. If the basic axioms and definitions are thus translated, then it is small wonder that the translation generalizes; take any probabilistic theorem, apply the above translation to it, and you are almost guaranteed to get a ranking theorem. This translation is obviously committed to negative ranks; therefore I always favored negative over positive ranks. However, the translation is not fool-proof; see, e.g., Spohn (1994a) for slight failures concerning conditional independence (between sets of variables) or Spohn (2005a) for slight differences concerning positive and non-negative instantial relevance. The issue is not completely cleared up.

Is there a deeper reason why this translation works so well? Yes, of course. The translation of products and quotients of probabilities suggests that negative ranks simply are the logarithm of probabilities (with respect to some base < 1). This does not seem to fit with the translation of sums of probabilities. But it does fit when the logarithmic base is taken to be some infinitesimal i (since for two positive reals $x \leq y$ $i^x + i^y = i^{x-j}$ for some infinitesimal j). That is, we may understand ranks as real orders of magnitude of non-standard probabilities. This is the basic reason for the pervasive analogy.

Does this mean that ranking epistemology simply reduces to non-standard Bayesianism? This may be one way to view the matter. However, I do not particularly like this perspective. Bayesian epistemology in terms of non-standard

reals is really non-standard. Even its great proponent, David Lewis, mentions the possibility only in passing (for the first time in 1980, p. 268). It is well known that both, non-standard analysis and its continuation as hyperfinite probability theory, have their intricacies of their own, and it is highly questionable from an epistemological point of view whether one should buy these intricacies. Moreover, even though this understanding of ranks is in principle feasible, it is nowhere worked out in detail. Such an elaboration should also explain the slight failures of the above translation. Hence, even formally the relation between ranks and non-standard probabilities is not fully clear. Finally, there are algebraic incoherencies. As long as the probabilistic law of additivity and the ranking law of disjunction are finitely restricted, there is no problem. However, it is very natural to conceive probability measures as σ-additive (although there is an argument about this point), whereas it is very natural to conceive of ranking functions as complete (as I have argued). This is a further disanalogy, which is not resolved by the suggested understanding of ranks.

All in all, I prefer to stick to the realm of standard reals. Ranking theory is a standard theory, and it should be compared to other standard theories. So, let us put the issue of hyperfinite probability theory to one side.

Let us instead pursue another line of thought. I have heavily emphasized that the fundamental point of ranking theory is to represent the statics and the dynamics of belief or of taking-to-be-true; it *is* the theory of belief. So, instead of inquiring the relation between ranks and probabilities we might as well ask the more familiar question about the relation between belief and probability.

This relation is well known to be problematic. One naive idea is that belief vaguely marks some threshold in probability, i.e., that A is believed iff its subjective probability is greater than $1 - \varepsilon$ for some small ε. But this will not do, as is hightlighted by the famous lottery paradox (see Kyburg 1961, p. 197, and Hempel (1962, pp. 163–166). According to this idea you may believe A and believe B, but fail to believe $A \,\&\, B$. However, this amounts to saying that you do not know the truth table of conjunction, i.e., that you have not grasped conjunction at all. So, this idea is a bad one, as almost all commentators to the lottery paradox agree. One might think then about more complicated relations between belief and probability, but I confess not to have seen any convincing one.

The simplest escape from the lottery paradox is, of course, to equate belief with probability 1. This proposal faces two further problems, though. First, it seems intuitively inadequate to equate belief with maximal certainty in probabilistic terms; beliefs need not be absolutely certain. Secondly, but this is only a theoretical version of the intuitive objection, only belief expansion makes sense according to this proposal, but no genuine belief revision. Once you assign probability 1 to a proposition, you can never get rid of it according to all rules of probabilistic change. This is obviously inadequate; of course, we can give up previous beliefs and easily do so all the time.

Jeffrey's radical probabilism (1991) is a radical way out. According to Jeffrey, all subjective probabilities are regular, and his generalized conditionalization provides a dynamics moving within regular probabilities. However, Jeffrey's picture and the

proposal of equating belief with probability 1 do not combine; then we would believe in nothing but the tautology. Jeffrey did not deny beliefs, but he indeed denied their relevance for epistemology; this is what the adjective 'radical' in effect signifies. He did not believe in any positive relation between belief and probability, and probability is all you need – a viable conclusion from the lottery paradox perhaps, though only as a last resort.

The point that probability theory cannot account for belief revision may apparently be dealt with by an expansion of the probabilistic point of view, namely by resorting to Popper measures. These take conditional probability as the basic notion, and thus probabilities conditional on propositions having absolute probability 0 may be well defined. That is, you may initially believe A, i.e., assign probability 1 to A, and still learn that $\overline{A}$, i.e., conditionalize w.r.t. $\overline{A}$, and thus move to posterior probabilities and even beliefs denying A. In this way, one can stick to the equation of belief with probability 1 and escape the above objection. Have we thus reached a stable position?

No, we have not. One point of Spohn (1986) was to rigorously show that AGM belief revision is just the qualitative counterpart of Popper measures. Conversely, this entails that the inability of AGM belief revision theory to model iterated belief revision, which I criticized in my (1988), holds for Popper measures as well. In fact, Harper (1976) was the first to note this problem vis à vis Popper measures, and thus I became aware of the problem and noticed the parallel.

Harper proposed quite a complicated solution to the problem that is, as far as I know, not well received; but it may be worth revisiting. My conclusion was a different one. If AGM belief revision theory is incomplete and has to be evolved into ranking theory, the probabilistic point of view needs likewise to get further expanded. We need something like probabilified ranks or ranked probabilities; it is only in terms of them that we can unrestrictedly explain iterated probabilistic change.

A ranking function associates with each rank a set of propositions having that rank. A ranked probability measure associates with each rank an ordinary probability measure. The precise definition is straightforward. Hence, I confined myself to mentioning the idea in my (1988, sect. 7); only in my (2005b) I took the trouble to explicitly introduce it. One should note, though, that as soon as one assumes the probability measures involved to be σ-additive, one again forces the ranks to be well-ordered (cf. Spohn 1986); this is why in my (2005b) only the probabilification of complete ranking functions is defined.

One may say that ranking theory thus ultimately reduces to probability theory. I find this misleading, however. What I have just sketched is rather a unification of probability and ranking theory; after all, we have employed genuine ranking ideas in order to complete the probabilistic point of view. The unification is indeed a powerful one; all the virtues of standard Bayesianism which I have shown to carry over to ranking theory hold for this unification as well. It provides a unified account of confirmation, of lawlikeness, even of causation. It appears to be a surprising, but most desirable wedding of Baconian and Pascalian probability. I shall continue on the topic in the next subsection.

The previous paragraphs again urge the issue of hyperfinite probability; ranked probabilities look even more like probabilities in terms of non-standard reals. However, I cannot say more than I already did; I recommend the issue for further investigation.[21] I should use the occasion for clarifying a possible confusion, though. McGee (1994, pp. 181ff.) showed that Popper measures correspond to non-standard probability measures in a specific way. Now, I have suggested that ranked probabilities do so as well. However, my (1986, 1988) together entail that ranked probabilities are more general than Popper measures. These three assertions do not fit together. Yet, the apparent conflict is easily dissolved. The correspondence proved by McGee is not a unique one. Different non-standard probability measures may correspond to the same Popper measure, just as different ranked probabilities may. Hence, if McGee says that the two approaches, Popper's and the non-standard one, "amount to the same thing" (p. 181), this is true only for the respects McGee is considering, i.e., w.r.t. conditional probabilities. It is not true for the wider perspective I am advocating here, i.e., w.r.t. probability dynamics.

Philosophical Aspects

The relation between belief and probability is not only a formal issue, it is philosophically deeply puzzling. It would be disturbing if there should be two (or more) unrelated ways of characterizing our doxastic states. We must somehow come to grips with their relation.

The nicest option would be *reductionism*, i.e., reducing one notion to the other. This can only mean reducing belief to probability. As we have seen, however, this option seems barred by the lottery paradox. Another option is *eliminativism* as most ably defended in Jeffrey's radical probabilism also mentioned above. This option is certainly viable and most elegant. Still, I find it deeply unsatisfactory; it is unacceptable that our talk of belief should merely be an excusable error ultimately to be eliminated. Thus, both versions of *monism* seem excluded.

Hence, we have to turn to *dualism*, and then *interactionism* may seem the most sensible position. Of course, everything depends on the precise form of interaction between belief and probability. In Spohn (2005b) I had an argument with Isaac Levi whom I there described as the champion of interactionism. My general experience, though, is that belief and probability are like oil and water; they do not mix easily. Quite a different type of interactionism is represented by Hild (t.a.) who has many interesting things to say about how ranking and probability theory mesh, indeed how heavily ranking ideas are implicitly used in statistical methodology. I do not have space to assess this type of interactionism.

[21] For quite a different way of relating probabilities and ranks appealing neither to infinitesimals nor to Popperian conditional probabilities see Giang and Shenoy (1999).

When the fate of interactionism is unclear one might hope to return to reductionism and thus to monism, not in the form of reducing belief to probability, but in the form of *reducing both to something third*. This may be hyperfinite probability, or it may be ranked probabilities as suggested above. However, as already indicated, I consider this to be at best a formal possibility with admittedly great formal power of unification. Philosophically, I am not convinced. It is intuitively simply inadequate to equate belief with (almost) maximal probabilistic certainty, i.e., with probability 1 (minus an infinitesimal), even if this does not amount to unrevisability within these unifications. This intuition has systematic counterparts. For centuries, the behavioral connection of subjective probabilities to gambling and betting has been taken to be fundamental; many hold that this connection provides the only explanation of subjective probabilities. This fundamental connection does not survive these unifications. According to them, I would have to be prepared to bet my life on my beliefs; but this is true only of very few of my many beliefs. So, there are grave frictions that should not be plastered by formal means.

In view of all this, I have always preferred *separatism*, at least *methodologically*. If monism and interactionism are problematic, then belief and probability should be studied as two separate fields of interest. I sense the harshness of this position; this is why I am recommending it so far only as a methodological one and remain unsure about its ultimate status. However, the harshness is softened by the formal parallel which I have extensively exploited and which allows formal unification. Thus, separatism in effect amounts to *parallelism*, at least if belief is studied in ranking terms. Indeed, the effectiveness of the parallel sometimes strikes me as a pre-established harmony.

Thus, another moral to be drawn may perhaps be *structuralism*, i.e., the search for common structures. This is a strategy I find most clearly displayed in Halpern (2003). He starts with a very weak structure of degrees of belief that he calls plausibility measures and then discusses various conditions on those degrees that allow useful strengthenings of that structure such as a theory of conditioning, a theory of independence, a theory of expectation and integration, and so forth. Both, ranking and probability theory, but not only they are specializations of that structure and its various strengthenings. Without doubt, this is a most instructive procedure. Structuralism would moreover suggest that it is only those structures and not their specific realizations that matter. Halpern does not explicitly endorse this, and I think one should withstand it. For instance, one would thereby miss the essential purpose for which ranking theory was designed, namely the theory of belief. For this purpose, no less and no more than the ranking structure is required.

Hence, let me further pursue, in the spirit of methodological separatism, the philosophical comparison between ranks and standard probabilities. I have already emphasized the areas in which the formal parallel also makes substantial sense: inductive inference, confirmation, causation, etc. Let us now focus on three actual or apparent substantial dissimilarities, which in one or the other way concern the issue what our doxastic states have to do with reality.

The first aspect of this issue is the *truth connection*; ranks are related to truth in a way in which probabilities are not. This is the old point all over again.

Ranks represent beliefs that are true or false, whereas subjective probabilities do not represent beliefs and may be assessed in various ways, as well-informed, as reasonable, but never as true or false. Degrees of belief may perhaps conform to degrees of truthlikeness; however, it is not clear in the first place whether degrees of truthlikeness behave like probabilities (cf. Oddie 2001). Or degrees of belief may conform to what Joyce (1998) calls the norm of gradational accuracy from which he proceeds with an interesting argument to the effect that degrees of belief then have to behave like probabilities.[22] Such ideas are at best a weak substitute, however; they never yield an application of truth in probability theory as we have it in ranking theory.

This is a clear point in favor of ranking theory. And it is rich of consequences. It means that ranking theory, in contrast to probability theory, is able to connect up with traditional epistemology. For instance, Plantinga (1993, chs. 6 and 7) despairs of finding insights in Bayesianism he can use and dismisses it, too swiftly I find. This would have been different with ranking theory. The reason why ranking theory is connectible is obvious. Traditional epistemology is interested in knowledge, a category entirely foreign to probability theory; knowledge, roughly, is justified true belief and thus analyzed by notions within the domain of ranking theory. Moreover, the notion of justification has become particularly contested in traditional epistemology; one focal issue was then to give an account of the truth-conduciveness of reasons, again notions within the domain of ranking theory.

I am not claiming actual epistemological progress here. But I do claim an advantage of ranking over probability theory, I do claim that traditional epistemology finds in ranking theory adequate formal means for discussing its issues, and using such means is something I generally recommend as a formal philosopher.

The second aspect is the *behavioral connection*. Our doxastic states make some actions rational and others irrational, and our theories have to say which. Here, probability theory seems to have a clear advantage. The associated behavioral theory is, of course, decision theory with its fundamental principle of maximizing conditional expected utility. The power of this theory need not be emphasized here. Is there anything comparable on offer for ranking theory?

This appears excluded, for the formal reason that there is a theory of integration and thus of expectation in probabilistic, but none in ranking terms; this is at least what I had thought all along. However, the issue has developed. There are various remarkable attempts of stating a decision theory in terms of non-probabilistic or non-additive representations of degrees of belief employing the more general Choquet theory of integration.[23] Indeed, there is also one especially for ranking theory. Giang and Shenoy (2000) translate the axiomatic treatment of utility as it is

[22] Cf., however, Maher's (2002) criticism of Joyce's argument.

[23] Economists inquired the issue; see, e.g., Gilboa (1987), Schmeidler (1989), Jaffray (1989), Sarin and Wakker (1992) for early contributions, and Wakker (2005) for a recent one. The AI side concurs; see, e.g., Dubois and Prade (1995), Brafman and Tennenholtz (2000), and Giang and Shenoy (2005).

given by Luce and Raiffa (1957, sect. 2.5) in terms of simple and compound lotteries directly into the ranking framework, thus developing a notion of utility fitting to this framework. These attempts doubtlessly deserve further scrutiny (cf. also Halpern 2003, ch. 5).

Let me raise, though, another point relating to this behavioral aspect. Linguistic behavior is unique to humans and a very special kind of behavior. Still, one may hope to cover it by decision theoretic means, too. Grice's intentional semantics employs a rudimentary decision theoretic analysis, and Lewis' (1969) theory of conventions uses game (and thus decision) theoretic methods in a very sophisticated way. However, even Lewis' account of coordination equilibria may be reduced to a qualitative theory (in Lewis (1975) he explicitly uses only qualitative terminology). In fact, the most primitive linguistic behavioral law is the disquotation principle: if *a* seriously and sincerely utters "*p*", then *a* believes that *p*.[24] The point is that these linguistic behavioral laws and in particular the disquotation principle is stated in terms of belief. There is no probabilistic version of the disquotation principle, and it is unclear what it could be. The close relation between belief and meaning is obvious and undoubted, though perhaps not fully understood in the philosophy of language. I am not suggesting that there is a linguistic pragmatics in terms of ranking functions; there is hardly anything.[25] I only want to point out that the standing of ranking theory concerning this behavioral aspect is at least promising.

There is a third and final aspect, again apparently speaking in favor of probability theory. We do not only make decisions with the help of our subjective probabilities, we also do *statistics*. That is, we find a lot of *relative frequencies* in the world, and they are closely related to probabilities. We need not discuss here the exact nature of this relation. Concerning objective probabilities, it is extensively discussed in the debate about frequentism, and concerning subjective probabilities it is presumably best captured in Reichenbach's principle postulating that our subjective probabilities should rationally converge to the observed relative frequencies. What is clear, in any case, is that in some way or other relative frequencies provide a strong anchoring of probabilities in reality from which the powerful and pervasive application of statistical methods derives. Subjective probabilities are not simply free-floating in our minds.

For many years I thought that this is another important aspect in which ranking theory is inferior to probability theory. Recently, though, I have become more optimistic. Not that there would be any statistics in ranking terms[26]; I do not see ranks related to relative frequencies. However, a corresponding role is played by the notion of *exception* and thus by absolute frequencies. In section "Objective ranks?",

[24] If *a* speaks a foreign language, the principle takes a more complicated, but obvious form. There is also a disquotation principle for the hearer, which, however, requires a careful exchange of the hearer's and the speaker's role.

[25] See in particular Merin (2006, appendix B) and (2008) whose relevance-based pragmatics yields interesting results in probabilistic as well as in ranking-theoretic terms.

[26] However, I had already mentioned that Hild (t.a.) finds a much closer connection of probabilities and ranks within statistical methodology.

I left the precise account of objectifiable ranking functions in the dark. If one studies that account more closely, though, one finds that these objectifiable ranking functions, or indeed the laws as I have indicated them in section "Objective ranks?", are exception or fault counting functions. The rank assigned to some possible world by such a ranking function is just the number of exceptions from the law embodied in this function that occur in this world.

This is a dim remark so far, and here is not the place to elaborate on it. Still, I find the opposition of exceptions and relative frequencies appealing. Often, we take a type of phenomenon as more or less frequent, and then we apply our sophisticated statistical methodology to it. Equally often, we try to cover a type of phenomenon by a deterministic law, we find exceptions, we try to improve our law, we take recourse to a usually implicit ceteris paribus condition, etc. As far as I know, the methodology of the latter perspective is less sophisticated. Indeed, there is little theory. Mill's method of relevant variables, e.g., is certainly an old and famous attempt to such a theory (cf. its reconstruction in Cohen 1977, ch. 13). Still, both perspectives, the statistical and the deterministic one, are very familiar to us. What I am suggesting is that the deterministic perspective can be thoroughly described in terms of ranking theory.[27]

It would moreover be most interesting to attend to the vague borderline. Somewhere, we switch from one to the other perspective, from exceptions to small relative frequencies or the other way around. I am not aware of any study of this borderline, but I am sure it is worth getting inquired. It may have the potential of also illuminating the relation of belief and probability, the deterministic and the statistical attitude.

All these broad implications are involved in a comparison of ranks and probabilities. I would find it rather confusing to artificially combine them in some unified theory, be it hyperfinite or ranked probabilities. It is more illuminating to keep them separate. Also, I did not want to argue for any preference. I wanted to present the rich field of comparison in which both theories can show their great, though partially diverging virtues. There should be no doubt, however, that the driving force behind all these considerations is the formal *parallelism* which I have extensively used in section "The theory" (pp. 305ff) and explained in section "Formal aspects" (pp. 328ff).

Further Comparisons

Let me close the paper with a number of brief comparative remarks about alternative accounts subsumable under the vague label 'Baconian probability'. I have already

[27] I attempted to substantiate this suggestion with my account of strict and ceteris paribus laws in Spohn (2002) and with my translation of de Finetti's representation theorem into ranking theory in Spohn (2005a). (New addendum: For the most recent ranking-theoretic account of ceteris paribus laws see Spohn (2014).)

made a lot of such remarks *en passant*, but it may be useful to have them collected. I shall distinguish between the earlier and usually more philosophical contributions on the one hand and the more recent, often more technical contributions from the computer science side on the other hand. The borderline is certainly fuzzy, and I certainly do not want to erect boundaries. Still, the centuries old tendency of specialization and of transferring problems from philosophy to special fields may be clearly observed here as well.

Earlier and Philosophical Literature

It is perhaps appropriate to start with L. Jonathan Cohen, the inventor of the label. In particular his (1977) is an impressive document of dualism, indeed separatism concerning degrees of provability and degrees of probability or inductive (Baconian) and Pascalian probability. His work is, as far as I know, the first explicit and powerful articulation of the attitude I have taken here as well.[28]

However, his functions of inductive support are rather a preform of my ranking functions. His inductive supports correspond to my positive ranks. Cohen clearly endorsed the law of conjunction for positive ranks; see his (1970, pp. 21f. and p. 63). He also endorsed the law of negation; but he noticed its importance only in his (1977, pp. 177ff.), whereas in his (1970) it is well hidden as theorem 306 on p. 226. His presentation is a bit imperspicuous, though, since he is somehow attached to the idea that $\Box^i$, i.e., having an inductive support $\geq i$, behaves like iterable S4-necessity and since he even brings in first-order predicate calculus.

Moreover, Cohen is explicit on the relationality of inductive support; it is a two-place function relating evidence and hypothesis. Hence, one might expect to find a true account of conditionality. This, however, is not so. His conditionals behave like strict implication[29], a feature Lewis (1973, sect. 1.2–3) has already warned against. Moreover, Cohen discusses only laws of support with fixed evidence – with one exception, the consequence principle, as he calls it (1970, p. 62). Translated into my notation it says for a positive ranking function π that

$$\pi(C \mid A) \geq \pi(C \mid B) \text{ if } A \subseteq B, \tag{17.15}$$

which is clearly not a theorem of ranking theory. These remarks sufficiently indicate that the aspect so crucial for ranking functions is scarcely and wrongly developed in Cohen's work.

The first clear articulation of the basic Baconian structure is found, however, not in Cohen's work, but in Shackle (1949, 1969). His functions of potential surprise

[28] I must confess, though, that I had not yet noticed his work when I basically fixed my ideas on ranking functions in 1983.

[29] This is particularly obvious from Cohen (1970, p. 219, def. 5).

clearly correspond to my negative ranking functions; axiom (17.9) in (1969, p. 81) is the law of negation, and axiom (17.4) and/or (17.6) in (1969, p. 90) express the law of disjunction. At least informally, Shackle also recognizes the duality of positive and negative ranks. He is explicit that potential surprise expresses certainty of wrongness, i.e., disbelief, and that there is conversely certainty of rightness (1969, p. 74).

His general attitude, however, is not so decidedly dualistic as that of Cohen. His concern is rather a general account of uncertainty, and he insists that probability does not exhaust uncertainty. Probability is an appropriate uncertainty measure only if uncertainty is 'distributional', whereas potential surprise accounts for 'non-distributional' uncertainty. So, he also ends up with an antagonistic structure; but the intention was to develop two special cases of a general theory.

It is most interesting to see how hard Shackle struggles with an appropriate law of conjunction for negative ranks. The first version of his axiom 7 (1969, p. 80) claims, in our terminology, that

$$\kappa(A \cap B) = \max\{\kappa(A), \kappa(B)\}. \tag{17.16}$$

He accepts the criticism this axiom has met, and changes it into a second version (1969, p. 83), which I find must be translated into

$$\kappa(B) = \max\{\kappa(A), \kappa(B \mid A)\} \tag{17.17}$$

(and is hence no law of conjunction at all). He continues that it would be fallacious to infer that

$$\kappa(A \cap B) = \min[\max\{\kappa(A), \kappa(B \mid A)\}, \max\{\kappa(B), \kappa(A \mid B)\}]. \tag{17.18}$$

In (1969, ch. 24) he is remarkably modern in discussing "expectation of change of own expectations". I interpret his formula (i) on p. 199 as slightly deviating from the second version of his axiom 7 in claiming that

$$\kappa(A \cap B) = \max\{\kappa(A), \kappa(B \mid A)\}. \tag{17.19}$$

And on pp. 204f. he even considers, and rejects (for no convincing reason), the equation

$$\kappa(A \cap B) = \kappa(A) + \kappa(B \mid A), \tag{17.20}$$

i.e., our law of conjunction for negative ranks. In all these discussions, conditional degrees of potential surprise appear to be an unexplained primitive notion. So, Shackle may have been here on the verge of getting things right. On the whole, though, it seems fair to say that his struggle has not led to a clear result.

Isaac Levi has always pointed to this pioneering achievement of Shackle, and he has made his own use of it. In a way he did not develop Shackle's functions

of potential surprise; he just stuck to the laws of negation and of disjunction for negative ranks. In particular, there is no hint of any notion of conditionalization. This is not to say that his epistemology is poorer than the one I have. Rather, he finds a place for Shackle's functions in his elaborated doxastic decision theory, more precisely, in his account of belief expansion. He adds a separate account of belief contraction, and with the help of what is called Levi's identity he can thus deal with every kind of belief change. He may even claim to come to grips with iterated change.[30] One may thus sense that his edifice is at cross-purposes with mine.

A fair comparison is hence a larger affair. I have tried to give it in Spohn (2005b). Let me only mention one divergence specifically related to ranking functions. Since Levi considers ranking functions as basically identical with Shackle's functions of potential surprise and since he sees the latter's role in expansion, he continuously brings ranking functions into the same restricted perspective. I find this inadequate. I rather see the very same structure at work at expansions as well as at contractions, namely the structure of ranks. Insofar I do not see any need of giving the two kinds of belief change an entirely different treatment.

This brings me to the next comparison, with AGM belief revision theory (cf. e.g., Gärdenfors 1988). I have already explained that I came to think of ranking theory as a direct response to the challenge of iterated belief revision for AGM belief revision theory, and I have explained how $A \to x$-conditionalization for ranks unifies and generalizes AGM expansion, revision, and contraction. One may wonder how that challenge was taken up within the AGM discussion. With a plethora of proposals (see Rott 2008), that partially ventilated ideas that I thought to have effectively criticized already in Spohn (1988) and that do not find agreement, as far as I see, with the exception of Darwiche and Pearl (1997). As mentioned, Hild and Spohn (2008) gives a complete axiomatization of iterated contraction. Whether it finds wider acceptance remains to be seen.

By no means, though, one should underestimate the richness of the AGM discussion, of which, e.g., Rott (2001) or Hanson (1999) give a good impression. A pertinent point is that ranking theory generalizes and thus simply sides with the standard postulates for revision and contraction (i.e., (K*1-8) and (K^-1-8) in Gärdenfors 1988, pp. 54–56 and 61–64). The ensuing discussion has shown that these postulates are not beyond criticism and that many alternatives are worth discussing (cf., e.g., Rott 2001, pp. 103ff., who lists three alternatives of K*7, nine of K*8, six of K^-7, and ten of K^-8). I confess I would not know how to modify ranking theory in order to do justice to such alternatives. Hence, a fuller comparison with AGM belief revision theory would have to advance a defense of the standard postulates against the criticisms related with the alternatives.

The point is, of course, relevant in the debate with Levi, too. He prefers what he calls mild contraction to standard AGM contraction that can be represented in

[30]Many aspects of his epistemology are already found in Levi (1967). The most recent statement is given in Levi (2004), where one also gets a good idea of the development of his thought.

ranking theory only as a form of iterated contraction. Again, one would have to discuss whether this representation is acceptable.

It is worth mentioning that the origins of AGM belief revision theory clearly lie in conditional logic. Gärdenfors' (1978) epistemic semantics for conditionals was a response to the somewhat unearthly similarity spheres semantics for counterfactuals in Lewis (1973), and via the so-called Ramsey test Gärdenfors' interest more and more shifted from belief in conditionals to conditional beliefs and thus to the dynamics of belief. Hence, one finds a great similarity in the formal structures of conditional logic and belief revision theory. In particular, Lewis' similarity spheres correspond to Gärdenfors' entrenchment relations (1988, ch. 4). In a nutshell, then, the progress of ranking theory over Lewis' counterfactual logic lies in proceeding from an ordering of counterfactuality (as represented by Lewis' nested similarity spheres) to a cardinal grading of disbelief (as embodied in negative ranking functions).[31]

Indeed, the origins reach back farther. Conditional logic also has a history, the earlier one being somewhat indeterminate. However, the idea of having an ordering of levels of counterfactuality or of far-fetchedness of hypotheses is explicitly found already in Rescher (1964). If κ is a positive ranking function taking only finitely many values $0, x_1, \ldots, x_m, \infty$, then $\kappa^{-1}(\infty), \kappa^{-1}(x_m), \ldots, \kappa^{-1}(x_1), \kappa^{-1}(0)$ is just a family of modal categories $M_0, \ldots, M_n$ $(n = m + 2)$, as Rescher (1964, pp. 47–50) describes it. His procedure on pp. 49 f. for generating modal categories makes them closed under conjunction; this is our law of conjunction for positive ranks. And he observes on p. 47 that all the negations of sentences in modal categories up to M_{n-1} must be in $M_n = \kappa^{-1}(0)$; this is our law of negation.

To resume, I cannot find an equivalent to the ranking account of conditionalization in all this literature. However, the philosophical fruits I have depicted in section "The theory", pp. 305ff., and also in section "Philosophical aspects", pp. 331ff., sprang from this account. Therefore, I am wondering to which extent this literature can offer similar fruits, and for all I know the answer tends to be negative.

More Recent Computer Science Literature

In view of the exploding computer science literature on uncertainty since the 80s even the brief remarks in the previous subsection on the earlier times were disproportionate. However, it is important, I think, not to forget about the origins. My comparative remarks concerning the more recent literature must hence be even more cursory. This is no neglect, though, since Halpern (2003), in book length, provides comprehensive comparisons of the various approaches with an emphasis on those aspects (conditionalization, independence, etc.) that I take to be important, too. Some rather general remarks must do instead and may nevertheless be illuminating.

[31] For my ideas how to treat conditionals in ranking-theoretic terms see Spohn (2015).

In the computer science literature, ranking theory is usually subsumed under the heading "uncertainty" and "degrees of belief". This is not wrong. After all, ranks are degrees, and if (absolute) certainty is equated with unrevisability, revisable beliefs are uncertain beliefs. Still, the subsumption is also misleading. My concern was *not* to represent uncertainty and to ventilate alternative models of doing so. Thus stated, this would have been an enterprise with too little guidance. My concern was exclusively to statically and dynamically represent *ungraded* belief, and my observation was that this necessarily leads to the ranking structure. If this is so, then, as I have emphasized, all the philosophical benefits of having a successful representation of ungraded belief are conferred to ranking theory. By contrast, if one starts modeling degrees of uncertainty, it is always an issue (raised, for instance, by the lottery paradox vis à vis probability) to which extent such a model adequately captures belief and its dynamics. So, this is a principled feature that sets ranking theory apart from the entire uncertainty literature.

The revisability of beliefs was directly studied in computer science under headings like "default logic" or "nonmonotonic reasoning". This is another large and natural field of comparison for ranking theory. However, let me cut things short. The relation between belief revision theory and nonmonotonic reasoning is meticulously investigated by Rott (2001). He proved far-reaching equivalences between many variants on both sides. This is highly illuminating. At the same time, however, it is a general indication that the concerns that led me to develop AGM belief revision theory into ranking theory are not well addressed in these areas of AI. Of course, such lump-sum statements must be taken with caution.

The uncertainty literature has observed many times that the field of nonmonotonic reasoning is within its reach. Among many others, Pearl (1988, ch. 10) has investigated the point from the probabilistic side, and Halpern (2003, ch. 8) has summarized it from his more comprehensive perspective. This direction of inquiry is obviously feasible, but the reverse line of thought of deriving kinds of uncertainty degrees from kinds of nonmonotonic reasoning is less clear (though the results in Hild and Spohn (2008) about the measurement of ranks with via iterated contractions may be a step in the reverse direction).

So, let me return to accounts of uncertainty in a bit more detail, and let me take up *possibility theory* first. It originates from Zadeh (1978), i.e. from fuzzy set theory and hence from a theory of vagueness. Its elaboration in the book by Dubois and Prade (1988) and many further papers shows its wide applicability, but never denies its origin. So, it should at least be mentioned that philosophical accounts of vagueness (cf., e.g., Williamson 1994) have nothing much to do with fuzzy logic. If one abstracts from this interpretation, though, possibility theory is formally very similar to ranking theory. If *Poss* is a possibility measure, then the basic laws are:

$$Poss\,(\varnothing) = 0, Poss(W) = 1, \quad \text{and } Poss\,(A \cup B) = \max\{Poss(A), Poss(B)\}\,. \tag{17.21}$$

So far, the difference is merely one of scale. Full possibility 1 is negative rank 0, (im)possibility 0 is negative rank ∞, and translating the scales translates the

characteristic axiom of possibility theory into the law of disjunction for negative ranks. Indeed, Dubois and Prade often describe their degrees of possibility in such a way that this translation fits not only formally, but also materially.

Hence, the key issue is again how conditionalization is treated within possibility theory. There is some uncertainty. First, there is the motive that also dominated Shackle's account of the functions of potential surprise, namely to keep possibility theory as an ordinal theory where degrees of possibility have no arithmetical meaning. Then the idea is to stipulate that

$$Poss\,(A \cap B) = \min\{Poss(A), Poss\,(B \mid A)\} = \min\{Poss(B), Poss\,(A \mid B)\}. \tag{17.22}$$

This is just Shackle's proposal (17.19). Hisdal (1978) proposed to go beyond (17.19) just by turning (17.22) into a definition of conditional possibility by additionally assuming that conditionally things should be as possible as possible, i.e., by defining $Poss(B \mid A)$ as the maximal degree of possibility that makes (17.22) true:

$$Poss\,(B \mid A) = \left\{ \begin{array}{l} P\,(A \cap B)\,, \text{ if } Poss\,(A \cap B) < Poss(A) \\ 1, \text{ if } Poss\,(A \cap B) = Poss(A) \end{array} \right\}. \tag{17.23}$$

Halpern (2003, Proposition 3.9.2, Theorem 4.4.5, and Corollary 4.5.8) entails that Bayesian net theory works also in terms of conditional possibility thus defined. Many things, though, do not work well. It is plausible that $Poss(B \mid A)$ is between the extremes 1 and $Poss(A \cap B)$. However, (17.23) implies that it can take only those extremes. This is unintelligible. (17.22) implies that, if neither $Poss(B \mid A)$ nor $Poss(A \mid B)$ is 1, they are equal, a strange symmetry. And so on. Such unacceptable consequences spread through the entire architecture.

However, there is a second way to introduce conditional possibilities (cf., e.g., Dubois and Prade 1998, p. 206), namely by taking numerical degrees of possibility seriously and defining

$$Poss\,(B \,||\, A) = Poss\,(A \cap B)\,/Poss(A). \tag{17.24}$$

This looks much better. Indeed, if we define $\kappa(A) = \log Poss(A)$, the logarithm taken w.r.t. some positive base < 1, then κ is a negative ranking function such that also $\kappa(B \mid A) = \log Poss(B \,||\, A)$. Hence, (17.24) renders possibility and ranking theory isomorphic, and all the philosophical benefits may be gained in either terms. Still, there remain interpretational differences. If we are really up to degrees of belief and disbelief, then the ranking scale is certainly more natural; this is particularly clear when we look at the possibilistic analogue to two-sided ranking functions. My remarks about objectifiable ranking functions as fault counting functions would make no sense for a possibilistic scale. And so on. Finally, one must be aware that the philosophical benefits resulted from adequately representing *belief*. Hence, it is doubtful whether the formal structure suffices to maintain the benefits for alternative interpretations of possibility theory.

Let me turn to some remarks about (*Dempster-Shafer*) *DS belief functions*. Shafer (1976) built on Dempster's ideas for developing a general theory of evidence. He saw clearly that his theory covered all known conceptions of degrees of belief. This, and its computational manageability, explains its enormous impact. However, before entering any formal comparisons the first argument that should be settled is a philosophical one about the nature of evidence. There is the DS theory of evidence, and there is a large philosophical literature on observation and confirmation, Bayesianism being its dominant formal expression. I have explained why ranking theory and its account of reasons is a member of this family, too. Of course, this argument cannot even be started here. My impression, though, is that it is still insufficiently fought out, certainly hampered by disciplinary boundaries.

In any case, it is to be expected that DS belief functions and ranking functions are interpretationally at cross-purposes. This is particularly clear from the fact that negative ranking functions, like possibility measures or Shackle's functions of potential surprise, are formally a special case of DS belief functions; they are *consonant* belief functions as introduced in Shafer (1976, ch. 10). There, p. 219, Shafer says that consonant belief functions "are distinguished by their failure to betray even a hint of conflict in the evidence"; they "can be described as 'pointing in a single direction'." From the perspective of Shafer's theory of evidence this may be an adequate characterization. As a description of ranking functions, however, it does not make any sense whatsoever. This emphasizes that the intended interpretations diverge completely.

Even formally things do not fit together. We saw that the virtues of ranking theory depend on the specific behavior of conditional ranks. This does not generalize to DS belief functions. There is again an uncertainty how to conditionalize DS belief functions; there are two main variants (cf. Halpern 2003, p. 103 and p. 132, which I use as my reference book in the sequel). The central tool of Shafer's theory of evidence is the rule of combination proposed by Dempster (1967); it is supposed to drive the dynamics of DS belief functions. Combination with certain evidence is identical with one of the two variants of conditionalization (cf. Halpern 2003, p. 94). According to Shafer, other uncertain evidence is also to be processed by this rule. One might think, though, instead to handle it with Jeffrey's generalized conditionalization, which is indeed definable for both kinds of conditional belief functions (cf. Halpern 2003, p. 107). However, both kinds of Jeffrey conditionalization diverge from the rule of combination (cf. Halpern 2003, p. 107 and p. 114).

Indeed, this was my argument in Spohn (1990, p. 156) against formally equating ranking functions with consonant belief functions: Ranking dynamics is driven by a ranking analogue to Jeffrey conditionalization, but it cannot be copied by the rule of combination since the corresponding combinations move outside the realm of consonant belief functions. And, as I may add now, it does not help to let the dynamics of DS belief functions be driven by Jeffrey conditionalization instead of the rule of combination: Consonant belief functions are not closed under Jeffrey

conditionalization as well, whereas ranking functions are thus closed.[32] I conclude that there is no formal subsumption of ranking functions under DS belief functions. Hence, their interpretations do not only actually diverge, they are bound to do so.

Smets' transferable belief model (cf., e.g., Smets 1998) proposes a still more general model for changing DS belief functions in terms of his so-called specializations. One should check whether it offers means for formally subsuming ranking functions under his model. Even if this would be possible, however, the interpretational concerns remain. Smets' specializations are so much wedded to Shafer's conception of evidence that any subsumption would appear artificial and accidental. The philosophical argument about the nature of evidence is even more pressing here.

A final remark: There is a bulk of literature treating doxastic uncertainty not in terms of a specific probability measure, but in terms of convex sets of probability measures. The basic idea behind this is that one's uncertainty is so deep that one is not even able to fix one's subjective probability. In this case, doxastic states may be described as sets of measures or in terms of probability intervals or in terms of lower and upper probabilities. Again, the multiple ways of elaborating this idea and their relations are well investigated (see again Halpern 2003). Indeed, DS belief functions, which provide a very general structure, emerges as generalizations of lower probabilities. Even they, though, do not necessarily transcend the probabilistic point of view, as Halpern (2003, p. 279) argues; DS belief functions are in a way tantamount to so-called inner measures. May we say, hence, that the alternative formal structures mentioned ultimately reduce to probabilism (liberalized in the way explained)? We may leave the issue open, though it is obvious that the liberal idea of uncertainty conceived as sets of subjective probabilities is, in substance, a further step away from the ideas determining ranking theory. Even if probabilism were successful in this way, as far as ranking theory is concerned we would only be thrown back to our comparative remarks in section "Ranks and probabilities", pp. 328ff.

We may therefore conclude that ranking theory is a strong independent pillar in that confusingly rich variety of theories found in the uncertainty literature. This conclusion is the only point of my sketchy comparative remarks. Of course, it is not to deny that the other theories serve other purposes well. It is obvious that we are still far from an all-purpose account of uncertainty or degrees of belief.

[32] Does this contradict the fact that ranking functions are equivalent to possibility measures (with their second kind of conditionalization), that possibility measures may be conceived as a special case of DS belief (or rather: plausibility) functions, and that Jeffrey conditionalization works for possibility measures as defined by Halpern (2003, p. 107)? No. The reason is that Jeffrey conditionalization for possibility measures is not a special case of Jeffrey conditionalization for DS belief functions in general. Cf. Halpern (2003, p. 107).

References

Bacon, F. (1620), *Novum Organum*.

Brafman, R. I., & Tennenholtz, M. (2000). An axiomatic treatment of three qualitative decision criteria. *Journal of the Association of Computing Machinery, 47*, 452–482.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: Chicago University Press.

Cohen, L. J. (1970). *The implications of induction*. London: Methuen.

Cohen, L. J. (1977). *The probable and the provable*. Oxford: Oxford University Press.

Cohen, L. J. (1980). Some historical remarks on the Baconian conception of probability. *Journal of the History of Ideas, 41*, 219–231.

Darwiche, A., & Pearl, J. (1997). On the logic of iterated belief revision. *Artificial Intelligence, 89*, 1–29.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society B, 41*, 1–31.

de Finetti, B. (1937). La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré, 7*; engl. translation: (1964) Foresight: Its logical laws, its subjective sources. In: H. E. Kyburg, Jr., H. E., & Smokler (Eds.), *Studies in subjective probability* (pp. 93–158). New York: Wiley.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics, 38*, 325–339.

Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B, 30*, 205–247.

Dubois, D., & Prade, H. (1988). *Possibility theory: An approach to computerized processing of uncertainty*. New York: Plenum Press.

Dubois, D., & Prade, H. (1995). *Possibility theory as basis for qualitative decision theory*. In: Proceedings of the 14th International Joint Conference on Artificial Iintelligence (IJCAI'95), Montreal, pp. 1925–1930.

Dubois, D., & Prade, H. (1998). Possibility theory: Qualitative and quantitative aspects. In D.M. Gabbay & P. Smets (Eds.), *Handbook of defeasible reasoning and uncertainty management systems* (Vol. 1) (pp. 169–226). Dordrecht: Kluwer.

Gabbay, D. M., et al. (Eds.). (1994). *Handbook of logic in artificial intelligence and logic programming, vol. 3, nonmonotonic reasoning and uncertainty reasoning*. Oxford: Oxford University Press.

Garber, D. (1980). Field and Jeffrey conditionalization. *Philosophy of Science, 47*, 142–145.

Gärdenfors, P. (1978). Conditionals and changes of belief. In I. Niiniluoto & R. Tuomela (Eds.), *The logic and epistemology of scientific change* (pp. 381–404). Amsterdam: North-Holland.

Gärdenfors, P. (1988). *Knowledge in flux*. Cambridge: MIT Press.

Geiger, D., & Pearl, J. (1990). On the logic of causal models. In R. D. Shachter, T. S. Levitt, J. Lemmer, & L. N. Kanal (Eds.), *Uncertainty in artificial intelligence 4* (pp. 3–14). Amsterdam: Elsevier.

Giang, P. G., & Shenoy, P. P. (1999). On transformations between probability and spohnian disbelief functions. In K. B. Laskey & H. Prade (Eds.), *Uncertainty in artificial intelligence* (Vol. 15, pp. 236–244). San Francisco: Morgan Kaufmann.

Giang, P. G., & Shenoy, P. P. (2000). A qualitative linear utility theory for Spohn's theory of epistemic beliefs. In C. Boutilier & M. Goldszmidt (Eds.), *Uncertainity in artificial intelligence* (Vol. 16, pp. 220–229). San Francisco: Morgan Kaufmann.

Giang, P. G., & Shenoy, P. P. (2005). Two axiomatic approaches to decision making using possibility theory. *European Journal of Operational Research, 162*, 450–467.

Gilboa, I. (1987). Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics, 16*, 65–88.

Goldszmidt, M., & Pearl, J. (1996). Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence, 84*, 57–112.

Hacking, I. (1975). *The emergence of probability*. Cambridge: Cambridge University Press.

Halpern, J. Y. (2003). *Reasoning about uncertainty*. Cambridge: MIT Press.
Hansson, S.O. (ed.) (1997), *Special Issue on Non-Prioritized Belief Revision. Theoria* 63, 1–134.
Hansson, S. O. (1999). *A textbook of belief dynamics. Theory change and database updating*. Dordrecht: Kluwer.
Harper, W. L. (1976). Rational belief change, popper functions and counterfactuals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. I, pp. 73–115). Dordrecht: Reidel.
Hempel, C. G. (1945). Studies in the logic of confirmation. *Mind, 54*, 1–26 + 97–121.
Hempel, C. G. (1962). Deductive-nomological vs. Statistical explanation. In H. Feigl & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science, vol. III, scientific explanation, space, and time* (pp. 98–169). Minneapolis: University of Minnesota Press.
Hild, M. (t.a.). *Introduction to induction: On the first principles of reasoning*, Manuscript.
Hild, M., & Spohn, W. (2008). The measurement of ranks and the laws of iterated contraction. *Artificial Intelligence, 172*, 1195–1218.
Hintikka, J. (1962). *Knowledge and belief*. Ithaca: Cornell University Press.
Hisdal, E. (1978). Conditional possibilities – independence and noninteractivity. *Fuzzy Sets and Systems, 1*, 283–297.
Hohenadel, S. (2013). *Efficient epistemic updates in rank-based belief networks*. Dissertation, University of Konstanz. See http://nbn-resolving.de/urn:nbn:de:bsz:352-250406
Huber, F. (2006). Ranking functions and rankings on languages. *Artificial Intelligence, 170*, 462–471.
Huber, F. (2007). The consistency argument for ranking functions. *Studia Logica, 86*, 299–329.
Hunter, D. (1991). Graphoids, semi-graphoids, and ordinal conditional functions. *International Journal of Approximate Reasoning, 5*, 489–504.
Jaffray, J.-Y. (1989). Linear utility theory for belief functions. *Operations Research Letters, 8*, 107–112.
Jeffrey, R. C. (1965). *The logic of decision.* Chicago: University of Chicago Press, 2nd ed. 1983.
Jeffrey, R. C. (1991). *Probability and the art of judgment*. Cambridge: Cambridge University Press.
Jensen, F. V. (2001). *Bayesian networks and decision graphs*. Berlin: Springer.
Joyce, J. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science, 65*, 575–603.
Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.
Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. I). New York: Academic.
Krüger, L., et al. (1987). *The probabilistic revolution. Vol. 1: ideas in history, Vol. 2: ideas in the sciences*. Cambridge: MIT Press.
Kyburg, H. E., Jr. (1961). *Probability and the logic of rational belief*. Middletown: Wesleyan University Press.
Lange, M. (2000). *Natural laws in scientific practice*. Oxford: Oxford University Press.
Levi, I. (1967). *Gambling with truth*. New York: A. A. Knopf.
Levi, I. (2004). *Mild contraction: Evaluating loss of information due to loss of belief*. Oxford: Oxford University Press.
Lewis, D. (1969). *Convention: A philosophical study*. Cambridge: Harvard University Press.
Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.
Lewis, D. (1975). Languages and language. In K. Gunderson (Ed.), *Minnesota studies in the philosophy of science* (Vol. VII, pp. 3–35). Minneapolis: University of Minnesota Press.
Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. II, pp. 263–293). Berkeley: University of California Press.
Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. New York: Wiley.
Maher, P. (2002). Joyce's argument for probabilism. *Philosophy of Science, 69*, 73–81.
McGee, V. (1994). Learning the impossible. In E. Eells & B. Skyrms (Eds.), *Probability and conditionals. Belief revision and rational decision* (pp. 179–199). Cambridge: Cambridge University Press.
Merin, A. (2006). *Decision theory of rhetoric*, book manuscript, to appear.

Merin, A. (2008). Relevance and reasons in probability and epistemic ranking theory. A study in cognitive economy. In: Forschungsberichte der DFG-Forschergruppe *Logik in der Philosophie* (Nr. 130). University of Konstanz.

Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems: Theory and algorithms*. New York: Wiley.

Oddie, G. (2001). Truthlikeness. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2001 Edition). http://plato.stanford.edu/archives/fall2001/entries/truthlikeness

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo: Morgan Kaufman.

Pearl, J. (2000). *Causality. Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Plantinga, A. (1993). *Warrant: The current debate*. Oxford: Oxford University Press.

Pollock, J. L. (1995). *Cognitive carpentry*. Cambridge: MIT Press.

Rescher, N. (1964). *Hypothetical reasoning*. Amsterdam: North-Holland.

Rescher, N. (1976). *Plausible reasoning*. Assen: Van Gorcum.

Rott, H. (2001). *Change, choice and inference: A study of belief revision and nonmonotonic reasoning*. Oxford: Oxford University Press.

Rott, H. (2008). Shifting priorities: Simple representations for twenty seven iterated theory change operators. In D. Makinson, J. Malinowski, & H. Wansing (Eds.), *Towards mathematical philosophy*. Dordrecht: Springer.

Sarin, R., & Wakker, P. P. (1992). A simple axiomatization of nonadditive expected utility. *Econometrica, 60*, 1255–1272.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica, 57*, 571–587.

Shackle, G. L. S. (1949). *Expectation in economics*. Cambridge: Cambridge University Press.

Shackle, G. L. S. (1969). *Decision, order and time in human affairs* (2nd ed.). Cambridge: Cambridge University Press.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton: Princeton University Press.

Shafer, G. (1978). Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences, 19*, 309–370.

Shenoy, P. P. (1991). On Spohn's rule for revision of beliefs. *International Journal of Approximate Reasoning, 5*, 149–181.

Smets, P. (1998). The Transferable Belief Model for Quantified Belief Representation. In D.M. Gabbay & P. Smets (Eds.), *Handbook of defeasible reasoning and uncertainty management systems* (Vol. 1) (pp. 267–301). Dordrecht: Kluwer.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Berlin: Springer, 2nd ed. 2000.

Spohn, W. (1976/78). *Grundlagen der Entscheidungstheorie*. Ph.D. thesis, University of Munich 1976, published: Kronberg/Ts.: Scriptor 1978, out of print, pdf-version at: http://www.uni-konstanz.de/FuF/Philo/Philosophie/philosophie/files/ge.buch.gesamt.pdf

Spohn, W. (1983). *Eine Theorie der Kausalität*, unpublished Habilitationsschrift, Universität München, pdf-version at: http://www.uni-konstanz.de/FuF/Philo/Philosophie/philosophie/files/habilitation.pdf

Spohn, W. (1986). The representation of Popper measures. *Topoi, 5*, 69–74.

Spohn, W. (1988). Ordinal conditional functions. A dynamic theory of epistemic states. In W. L. Harper & B. Skyrms (Eds.), *Causation in decision, belief change, and statistics* (Vol. II, pp. 105–134). Dordrecht: Kluwer.

Spohn, W. (1990). A general non-probabilistic theory of inductive reasoning. In R. D. Shachter, T. S. Levitt, J. Lemmer, & L. N. Kanal (Eds.), *Uncertainty in artificial intelligence* (Vol. 4, pp. 149–158). Amsterdam: Elsevier.

Spohn, W. (1991). A reason for explanation: Explanations provide stable reasons. In W. Spohn, B. C. van Fraassen, & B. Skyrms (Eds.), *Existence and explanation* (pp. 165–196). Dordrecht: Kluwer.

Spohn, W. (1993). Causal laws are objectifications of inductive schemes. In J. Dubucs (Ed.), *Philosophy of probability* (pp. 223–252). Dordrecht: Kluwer.

Spohn, W. (1994a). On the properties of conditional independence. In P. Humphreys (Ed.), *Patrick suppes: Scientific philosopher. Vol. 1: Probability and probabilistic causality* (pp. 173–194). Dordrecht: Kluwer.

Spohn, W. (1994b). On Reichenbach's principle of the common cause. In W. C. Salmon & G. Wolters (Eds.), *Logic, language, and the structure of scientific theories* (pp. 215–239). Pittsburgh: Pittsburgh University Press.

Spohn, W. (1999). Two coherence principles. *Erkenntnis, 50*, 155–175.

Spohn, W. (2001a). Vier Begründungsbegriffe. In T. Grundmann (Ed.), *Erkenntnistheorie. Positionen zwischen Tradition und Gegenwart* (pp. 33–52). Paderborn: Mentis.

Spohn, W. (2001b). Bayesian nets are all there is to causal dependence. In M. C. Galavotti, P. Suppes, & D. Costantini (Eds.), *Stochastic dependence and causality* (pp. 157–172). Stanford: CSLI Publications.

Spohn, W. (2002). Laws, ceteris paribus conditions, and the dynamics of belief. *Erkenntnis, 57*, 373–394; also in: Earman, J., Glymour, C., Mitchell, S. (Eds.). (2002). *Ceteris paribus laws* (pp. 97–118). Dordrecht: Kluwer.

Spohn, W. (2005a). Enumerative induction and lawlikeness. *Philosophy of Science, 72*, 164–187.

Spohn, W. (2005b). Isaac Levi's potentially surprising epistemological picture. In E. Olsson (Ed.), *Knowledge and inquiry: Essays on the pragmatism of Isaac Levi*. Cambridge: Cambridge University Press.

Spohn, W. (2006). Causation: An alternative. *British Journal for the Philosophy of Science, 57*, 93–119.

Spohn, W. (2012). *The laws of belief. Ranking theory and its philosophical applications*. Oxford: Oxford University Press.

Spohn, W. (2014). The epistemic account of ceteris paribus conditions. *European Journal for the Philosophy of Science, 4*(2014), 385–408.

Spohn, W. (2015). Conditionals: A unified ranking-theoretic perspective. *Philosophers' Imprint 15*(1)1–30; see: http://quod.lib.umich.edu/p/phimp/3521354.0015.001/

Studeny, M. (1989). Multiinformation and the problem of characterization of conditional independence relations. *Problems of Control and Information Theory, 18*, 3–16.

Wakker, P. P. (2005). Decision-foundations for properties of nonadditive measures: General state spaces or general outcome spaces. *Games and Economic Behavior, 50*, 107–125.

Williamson, T. (1994). *Vagueness*. London: Routledge.

Zadeh, L. A. (1975). Fuzzy logics and approximate reasoning. *Synthese, 30*, 407–428.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems, 1*, 3–28.

Part III
Decision Theory

Chapter 18
Introduction

Horacio Arló-Costa, Vincent F. Hendricks, and Johan van Benthem

The classical account of decision-making derives from seminal work done by Frank P. Ramsey (1926) and later on by Von Neumann and Morgenstern (1947). This work culminates later on with the influential account of Leonard Savage (1954), Ascombe and Aumann (1963) and de Finetti (1974). We can recapitulate here in a compact form the classical presentation by Von Neumann and Morgenstern. Define a lottery as follows: If $A_1, \ldots, A_m$ is a partition of the possible outcomes of an experiment with $\alpha_j = \Pr(A_j)$ for each j, then the lottery $(\alpha_1, \ldots, \alpha_m)$ awards prize z_j if A_j occur. We can assume that the choice of the partition events does not affect the lottery. We can then introduce some central axioms for preferences among lotteries.

Axiom 18.1 (Weak Order) There is a weak order, $\geq$, among lotteries such that $L_1 \geq L_2$ iff L_1 is not strictly preferred to L_2.

Then we have a second crucial axiom:

Axiom 18.2 (Independence) For each L, L_1, L_2, and $0 < a < 1$, $L_1 \geq L_2$ iff $aL_1 + (1 - a)L \geq aL_2 + (1 - a)L$.

Horacio Arló-Costa was deceased at the time of publication.

H. Arló-Costa (deceased)
Carnegie Mellon University, Pittsburgh, PA, USA

V.F. Hendricks (✉)
Center for Information and Bubble Studies, University of Copenhagen, Copenhagen, Denmark
e-mail: vincent@hum.ku.dk

J. van Benthem
University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Stanford University, Stanford, United States
e-mail: johan@science.uva.nl

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_18

A third Archimedean axiom is often introduced to guarantee that utilities are real valued. These axioms suffice to prove that there exists a utility over prizes U such that $(\alpha_1, \ldots, \alpha_m) > (\beta_1, \ldots, \beta_m)$ iff $\sum_{I=1,m} \alpha_i U(z_i) \leq \sum_{I=1,m} \beta_i U(z_i)$ This utility is unique up to positive affine transformation. Anscombe and Aumann (1963) introduced a fourth axiom designed to explicitly say that preferences among prizes did not vary with the state, or, as it is usually put that utilities are state-independent. Savage (1954) provided an alternative set of axioms for preferences among acts (functions from states to consequences) that do not rely on an auxiliary randomization. He then shows that these axioms lead to a unique probability and state-independent utility such that acts are ranked according to their expected utilities. Savage's postulates are consistent with VonNeumann and Morgenstern's three central axioms. The corresponding theory offers what we can consider as the standard received view on models of preference by maximizing expected utility.

Savage's account of decision-making has been challenged by a series of paradoxes and counterexamples. Two of these paradoxes occupy a central role in recent theorizing. The first was offered by Maurice Allais (1953), the second by Daniel Ellsberg (1961). We will discuss here as well two additional paradoxes: one initially suggested by the physicist William Newcomb (presented explicitly in a published form by Robert Nozick (1969)) and a paradox due to Teddy Seidenfeld which appears in the article reprinted here that he coauthored with Mark J. Schervish and Joseph B. Kadane.

We can focus first on Allais's conundrum. Consider three rewards, $r_1 = \$0$, $r_2 = \$1$ million, $r_3 = \$5$ million. Now consider the following lotteries: $L_1 = 1$ million for certain; $L_2 = (.01, .89, .10)$ with prizes $z_i = r_i$; $L_3 = (.90, .10)$ with prizes $z_1 = r_1$ and $z_2 = r_3$; $L_4 = (.89, .11)$ with prizes $z_1 = r_1$ and $z_2 = r_2$. Most people choose L_1 over L_2, and L_3 over L_4. If we assume that choices reveal the underlying preference it is easy to see that this violates the classical theory of expected utility. There are, nevertheless, many cognitive explanations for this behavior. For example, Ariel Rubinstein has suggested (Similarity and Decision Making under Risk (Is there a utility theory resolution to the Allais Paradox?) *Journal of Economic Theory*, 46, 145–153, 1988) that in the first choice subjects simply go for the sure thing and in the second choice the probabilities .11 and .10 are sufficiently similar but r_3 clearly dominates r_2. Although the use of the corresponding heuristic usually leads to choices that can be justified by the theory of expected utility in this case the use of the heuristics clashes with expected utility (see Seidenfeld's article for a detailed explanation). Many psychologists concluded that in situations of this type agents reliably commit certain cognitive errors and tried to construct a theory capable of predicting this type of behavior. Daniel Kahnemann and Amos Tversky proposed a theory of this sort in a paper initially published in 1979 in the journal *Econometrica* (see the reference in the article reprinted here). The theory in question is usually called Prospect Theory (PT). The axiomatic presentation of PT abandons the corresponding version of Independence. Philosophers had different types of reactions to Allais but in general they accepted that there are at least some versions of the paradox that constitute examples of systematic errors caused by the bias induced by the use of certain heuristics.

The paradox proposed by Daniel Ellsberg is quite different. We can present here the simplest version of the paradox. Urn A contains exactly 100 balls. 50 of these balls are solid black and the remaining 50 are solid white. Urn B contains exactly 100 balls. Each of these balls is either solid black or solid white, although the ratio of black balls to white balls is unknown. Consider now the following questions: How much would you be willing to pay for a ticket that pays $25 ($0) if the next random selection from Urn A results in black (white) ball? Repeat then the same question for Urn B. It is well known that subjects tend to offer higher maximum buying prices for urn A than for urn B. This indicates that subjects do not have identical probabilities for both urns (.5 for each color) as Savage's theory predicts. It is considerably less clear that this behavior has to be interpreted as some sort of error. Ellsberg himself saw this behavior as an indication that Savage's theory has to be amended to deal with situations where uncertainty and vague or imprecise probabilities are involved. One can perfectly think, for example, that probabilities remain indeterminate in the case of Urn B. There is a vast literature dealing with decisions under ambiguity that is reviewed in the article by Gilboa and Marinacci reprinted here. As Seidenfeld's article indicates there are two main choices: either embracing a theory that abandons Axiom 18.1 (Ordering) or alternatively embracing a theory that abandons Axiom 18.2 (Independence). Seidenfeld argues that abandoning Independence (a solution that is rather popular and that Ellsberg himself supported) has a costly price: it leads to a form of sequential incoherence. Seidenfeld's argument requires the use of axioms for sequential decision making that many have found controversial. Seidenfeld's article remains mainly concerned with normative solutions to the paradoxes. The article by Tversky and Kahnemann reprinted here intends to extend the initial version of prospect theory to the case of uncertainty as well. So, they think that the common choices elicited by Ellsberg constitute also an error. This implies having a conservative attitude regarding the normative status of standard decision theory that clearly clashes with the motivation and some of the central theoretical ideas that motivated Ellsberg's work.

Mark J. Schervish, Teddy Seidenfeld and Josheph B. Kadane question in their paper another central tenet of the standard theories of decision making: the assumption that utility has to be state-independent. They show via an ingenious example that the uniqueness of probability in standard representations is relative to the choice of what counts as a constant outcome. Moreover they prove an important result showing how to elicit a unique state-dependent utility. The result does not assume that there are prizes with constant value by introducing a new kind of hypothetical kind of act in which both the prize and the state of nature are determined by an auxiliary experiment.

Our final paradox is the one proposed by the physicist William Newcomb. Consider an opaque box and a transparent box. An agent may choose one or the other taking into account the following: The transparent box contains one thousand dollars that the agent plainly sees. The opaque box contains either nothing or one million dollars, depending on a prediction already made. The prediction was about the agent's choice. If the prediction was that the agent will take both boxes, then the opaque box is empty. On the other hand, if the prediction was that the agent will take

just the opaque box, then the opaque box contains a million dollars. The prediction is reliable. The agent knows all these features of his decision problem. So, we can depict the agent's options as follows:

	Prediction of one-boxing	Prediction of two-boxing
Take only one box	$\$M$	\$0
Take two boxes	$\$M + \$T$	$\$T$

It is clear that two-boxing dominates one-boxing (the prizes of two-boxing are better than the prizes of one-boxing in each state of nature). So, two-boxing is the adequate choice according to dominance. Given the hypothesis of reliability of prediction, a prediction of one-boxing has a high probability given one-boxing. Similarly, a prediction of two-boxing has a high probability given two-boxing. Therefore, one-boxing's expected utility exceeds two-boxing's expected utility. One-boxing is the rational choice according to the principle of expected-utility maximization. Should one be a one-boxer or a two-boxer?

The formula used to calculate expected utility in the second case is: $U(A) = \sum_{I=1,n} \rho(S_i \mid A)\, u(A, S_i)$, where A is an act, and S_i are relevant states of nature. Joyce and Gibbard argue that two-boxing can be rationalized if one appeals to a different way of calculating expected utility:

$$UC(A) = \sum_{I=1,n} \rho(A > S_i)\, u(A, S_i) = \sum_{I=1,n} \rho(S_i \setminus A)\, u(A, S_i),$$

where the connective ">" is a counterfactual conditional and $\rho(S_i \setminus A)$ indicates a deviant type of conditional probability (called *imaging*) proposed by David Lewis. This type of conditional probability can be articulated in a paradox-free manner such that the probability of a counterfactual conditional coincides with the corresponding conditional probability (that is that $\rho(A > S_i) = \rho(S_i \setminus A)$). Classical conditional probability cannot satisfy this equation on pain on triviality (this was shown also by Lewis in a seminal paper that appeared in 1976: Probabilities of Conditionals and Conditional Probabilities, *Philosophical Review* 85, 297–315). The bibliographical notes below contain various useful pointers to recent papers debating the tenability of the corresponding notion of causal decision theory.

Suggested Further Reading

A classical and still rather useful book presenting the received view in decision theory is Savage's influential and seminal book: *The Foundations of Statistics*, Dover Publications; 2 Revised edition (June 1, 1972). A slightly more accessible but pretty thorough textbook presentation of Savage's account and beyond is the monograph by David Kreps: *Notes on the Theory of Choice*, Westview Press (May 12, 1988).

The classical essay by Daniel Ellsberg introducing his now famous paradox continues to be a very important source in this area: "Risk, Ambiguity and the Savage Axioms," *Quarterly Journal of Economics*, 75: 643-669, 1961. Isaac Levi presented a unified normative view of both Allais and Ellsberg in: "The Paradoxes of Allais and Ellsberg," *Economics and Philosophy*, 2: 23-53, 1986.

This solution abandons ordering rather than independence unlike the solution proposed by Ellsberg himself. Solutions abandoning independence have been in general more popular. Some of the classical papers in this tradition appear in the bibliography of the paper by Gilboa and Marinacci reprinted here. Many of the responses to Allais have been descriptive rather than normative. Prospect theory is a classical type of response along these lines. We reprint here an article that intends to present a unified descriptive response to both Allais and Ellsberg. The reader can find an excellent, thorough and mathematically mature presentation of the contemporary state of the art in Prospect theory in a recent book published by Peter Wakker: *Prospect Theory for Risk and Ambiguity*, Cambridge University Press, Cambridge, 2011.

The debate about the foundations of causal decision theory is in a way still open. An excellent presentation of causal decision theory can be found in an important book by Jim Joyce: *The Foundations of Causal Decision Theory*, Cambridge Studies in Probability, Induction and Decision Theory, Cambridge, 2008. Joyce has also written an interesting piece answering challenges to causal decision theory: "Regret and Instability in Causal Decision Theory," forthcoming in the second volume of a special issue of *Synthese* devoted to the foundations of the decision sciences (eds.) Horacio Arlo-Costa and Jeffrey Helzner. This special issue contains as well an essay by Wolfgang Spohn that intends to articulate the main ideas of causal decision theory by appealing to techniques used in Bayesian networks: "Reversing 30 Years of Discussion: Why Causal Decision Theorists should be One-Box." This is a promising line of investigation that has also been considered preliminary in an insightful article by Christopher Meek and Clark Glymour: "Conditioning and Intervening", *British Journal for the Philosophy of Science* 45, 1001-1021, 1994. With regard to issues related to actual causation a useful collection is the book: *Causation and Counterfactuals*, edited by J. Collins, N. Hall and L.A. Paul, MIT Press, 2004.

Finally a paper by Joseph Halpern and Judea Pearl offers a definition of actual causes using structural equations to model counterfactuals, Halpern, J. Y. and Pearl, J. (2005) "Causes and explanations: a structural-model approach. Part I: Causes", *British Journal for Philosophy of Science* 56:4, 843-887. This paper articulates ideas about causation based on recent work on Bayesian networks and related formalisms. Current work in this area seems to point to an unification of causal decision theory and an account of causation based on Bayesian networks.

Chapter 19
Allais's Paradox

Leonard Savage

Introspection about certain hypothetical decision situations suggests that the sure-thing principle and, with it, the theory of utility are normatively unsatisfactory. Consider an example based on two decision situations each involving two gambles.[1]

Situation 1.	Choose between
Gamble 1.	\$500,000 with probability 1; and
Gamble 2.	\$2,500,000 with probability 0.1, \$500,000 with probability 0.89, status quo with probability 0.01.
Situation 2.	Choose between
Gamble 3.	\$500,000 with probability 0.11, status quo with probability 0.89; and
Gamble 4.	\$2,500,000 with probability 0.1, status quo with probability 0.9.

Many people prefer Gamble 1 to Gamble 2, because, speaking qualitatively, they do not find the chance of winning a *very* large fortune in place of receiving a large fortune outright adequate compensation for even a small risk of being left in the status quo. Many of the same people prefer Gamble 4 to Gamble 3; because, speaking qualitatively, the chance of winning is nearly the same in both gambles, so the one with the much larger prize seems preferable. But the intuitively acceptable pair of preferences, Gamble 1 preferred to Gamble 2 and Gamble 4 to Gamble 3, is not compatible with the utility concept or, equivalently, the sure-thing

Leonard Savage was deceased at the time of publication.

[1]This particular example is due to Allais (1953). Another interesting example was presented somewhat earlier by Georges Morlat (1954).

L. Savage (deceased)
Princeton University, New York, NY, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_19

principle. Indeed that pair of preferences implies the following inequalities for any hypothetical utility function.

$$U(\$500{,}000) > 0.1U(\$2{,}500{,}000) + 0.89U(\$500{,}000) + 0.1U(\$0),$$
$$0.1U(\$2{,}500{,}000) + 0.9U(\$0) > 0.11U(\$500{,}000) + 0.89U(\$0); \tag{3}$$

and these are obviously incompatible.

Examples[2] like the one cited do have a strong intuitive appeal; even if you do not personally feel a tendency to prefer Gamble 1 to Gamble 2 and simultaneously Gamble 4 to Gamble 3, I think that a few trials with other prizes and probabilities will provide you with an example appropriate to yourself.

If, after thorough deliberation, anyone maintains a pair of distinct preferences that are in conflict with the sure-thing principle, he must abandon, or modify, the principle; for that kind of discrepancy seems intolerable in a normative theory. Analogous circumstances forced D. Bernoulli to abandon the theory of mathematical expectation for that of utility (Bernoulli 1738). In general, a person who has tentatively accepted a normative theory must conscientiously study situations in which the theory seems to lead him astray; he must decide for each by reflection—deduction will typically be of little relevance—whether to retain his initial impression of the situation or to accept the implications of the theory for it.

To illustrate, let me record my own reactions to the example with which this heading was introduced. When the two situations were first presented, I immediately expressed preference for Gamble 1 as opposed to Gamble 2 and for Gamble 4 as opposed to Gamble 3, and I still feel an intuitive attraction to those preferences. But I have since accepted the following way of looking at the two situations, which amounts to repeated use of the sure-thing principle.

One way in which Gambles 1–4 could be realized is by a lottery with a hundred numbered tickets and with prizes according to the schedule shown in Table 19.1.

Table 19.1 Prizes in units of \$100,000 in a lottery realizing gambles 1–4

		Ticket number		
		1	2–11	12–100
Situation 1	Gamble 1	5	5	5
	Gamble 2	0	25	5
Situation 2	Gamble 3	5	5	0
	Gamble 4	0	25	0

[2] Allais has announced (but not yet published) an empirical investigation of the responses of prudent, educated people to such examples (Allais 1953).

Now, if one of the tickets numbered from 12 through 100 is drawn, it will not matter, in either situation, which gamble I choose. I therefore focus on the possibility that one of the tickets numbered from 1 through 11 will be drawn, in which case Situations 1 and 2 are exactly parallel. The subsidiary decision depends in both situations on whether I would sell an outright gift of $500,000 for a 10-to-1 chance to win $2,500,000—a conclusion that I think has a claim to universality, or objectivity. Finally, consulting my purely personal taste, I find that I would prefer the gift of $500,000 and, accordingly, that I prefer Gamble 1 to Gamble 2 and (contrary to my initial reaction) Gamble 3 to Gamble 4.

It seems to me that in reversing my preference between Gambles 3 and 4 I have corrected an error. There is, of course, an important sense in which preferences, being entirely subjective, cannot be in error; but in a different, more subtle sense they can be. Let me illustrate by a simple example containing no reference to uncertainty. A man buying a car for $2,134.56 is tempted to order it with a radio installed, which will bring the total price to $2,228.41, feeling that the difference is trifling. But, when he reflects that, if he already had the car, he certainly would not spend $93.85 for a radio for it, he realizes that he has made an error.

One thing that should be mentioned before this chapter is closed is that the law of diminishing marginal utility plays no fundamental role in the von Neumann-Morgenstern theory of utility, viewed either empirically or normatively. Therefore the possibility is left open that utility as a function of wealth may not be concave, at least in some intervals of wealth. Some economic-theoretical consequences of recognition of the possibility of non-concave segments of the utility function have been worked out by Friedman and myself (1948), and by Friedman alone (1953). The work of Friedman and myself on this point is criticized by Markowitz (1952).[3]

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axioms de l'école Americaine. *Econometrica, 21*, 503–546.

Archibald, G. C. (1959). Utility, risk, and linearity. *Journal of Political Economy, 67,* 437–450.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii academiae scientiarum imperialis Petropolitanae (for 1730 and 1731)*, 5, 175–192.

Centre National de Recherche Scientifique Fondements et applications de la théorie du risque en econometrie, Paris, Centre National de la Recherche Scientifique. (1954) *Report of an international econometric colloquium on risk, in which there was much discussion of utility*, held in Paris, May 12–17, 1952.

Friedman, M. (1953). Choice, chance, and personal distribution of income. *Journal of Political Economy, 61,* 277–290.

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy, 56*, 279–304, Reprinted, with a correction, in Stigler and Boulding (1952).

[3] See also Archibald (1959) and Hakansson (1970).

Hakansson, N. H. (1970). Friedman-Savage utility functions consistent with risk aversion. *Quarterly Journal of Economics, 84,* 472–487.
Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy, 60,* 151–158, Marshall, Alfred.
Stigler, G. J., & Boulding, K. E. (Eds.). (1952). *Readings in price theory*. Chicago: Richard D. Irwin.

Chapter 20
Decision Theory Without "Independence" or Without "Ordering"

What Is the Difference?

Teddy Seidenfeld

Introduction

It is a familiar argument that advocates accommodating the so-called paradoxes of decision theory by abandoning the "independence" postulate. After all, if we grant that choice reveals preference, the anomalous choice patterns of the Allais and Ellsberg problems (reviewed in section "Review of the Allais and Ellsberg "Paradoxes") violate postulate P2 ("sure thing") of Savage's (1954) system. The strategy of making room for new preference patterns by relaxing independence is adopted in each of the following works: Samuelson (1950), Kahneman and Tversky's "Prospect Theory" (1979), Allais (1979), Fishburn (1981), Chew and MacCrimmon (1979), McClennen (1983), and in closely argued essays by Machina (1982, 1983 [see the latter for an extensive bibliography]).

There is, however, a persistent underground movement that challenges instead the normative status of the "ordering" postulate for preference. Those whose theories evidence some misgivings about ordering include: Good (1952), C. A. B. Smith (1961), Levi (1974, 1980), Suppes (1974), Walley and Fine (1979), Wolfenson and Fine (1982), and Schick (1984). And abandoning ordering is a strategy that has been used to resolve group decision problems. For this see Savage (1954, section 7.2) and Kadane and Sedransk (1980), and see Kadane (1986) for an application to clinical trials. "Regret" models also involve a failure of ordering since choice-with-regret does not satisfy Sen's (1977) principle of "independence of irrelevant alternatives": Savage (1954, section 13.5), Bell and Raiffa (1979) and Bell (1982), Loomes and Sugden (1982), and Fishburn (1983) discuss regret.

T. Seidenfeld (✉)
Departments of Philosophy and Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: teddy@stat.cmu.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_20

Expected Utility for Simple Lotteries – A Review

For ease of exposition, let us adopt an axiomatization similar to the von Neumann and Morgenstern (1947) theory, as condensed by Jensen (1967). Let R be a set of β-many rewards (or payoffs), $R = (r_\alpha\colon \alpha \leq \beta)$. In the spirit of Debreu's (1959, chapter 4) presentation, we can think of R as including (infinitely divisible) monetary rewards. A (simple) lottery over R is a probability measure P on R with the added requirement that $P(X) = 1$ for some *finite* subset of rewards.

Lotteries are individuated according to the following (0th) *reduction postulate*: Let L_1, L_2 be two lotteries with probabilities P_1, P_2 and let $R_n = (r_1, \ldots, r_n)$ be the finite set of the union of the payoffs under these two lotteries. A convex combination, $\alpha L_1 + (1-\alpha)L_2$ $(0 \leq \alpha \leq 1)$, of the two lotteries is again a lottery with probability measure $\alpha P_1 + (1-\alpha)P_2$ over R_n. Thus, the set of lotteries is a mixture set $M(R)$ in the sense of Herstein and Milnor (1953).

Three postulates comprise expected utility theory:

(1) An ordering requirement: preference, $\precsim$, a relation over $M \times M$, is a weak-order. That is, $\precsim$ is reflexive, transitive, and all pairs of lotteries are comparable under $\precsim$. (Strict preference, $\prec$, and indifference, $\sim$, are defined relations.)

(2) An Archimedean requirement: If $L_1 \prec L_2$ and $L_2 \prec L_3$, there is a nontrivial convex combination of L_1 and L_3 strictly preferred (and another combination strictly dispreferred) to L_2. That is, there exist

$$0 < \alpha, \beta < 1 \text{ with } \alpha L_1 + (1-\alpha)\, L_3 \prec L_2 \text{ and } L_2 \prec \beta L_1 + (1-\beta)\, L_3.$$

The point in assuming that R includes (infinitely divisible) monetary payoffs is made clear by the additional stipulation that each lottery in M carries a sure-dollar equivalent (under $\precsim$):

$$\forall L \in \mathrm{M}\ \ \exists \$x \in \mathrm{R}\ \ \ (L \sim L_{\$x})\,, \qquad (*)$$

where $L_{\$x}$ is a degenerate lottery having only one prize, $\$x$. Then principle (2) deserves its title for, with (*) and the added stipulation that more is (strictly) better when it comes to money, (1) and (2) entail a real-valued utility representation for $\precsim$ (continuous in \$).[1] To simplify still further, let us restrict attention to lotteries with none but monetary payoffs.

[1] By assuming (*), we fix it that $M/\sim$ (no longer assumed to be a mixture set) has a countable dense subset in the $<$−order on $M/\sim$, e.g., the rational-valued sure-dollar equivalents. Then our first two postulates ensure a real-valued utility u on M with the property that $L_1 \prec L_2$ if and only if $u(L_1) < u(L_2)$. The point is that, without "independence," the usual Archimedean axiom is neither necessary nor sufficient for a real valued utility. See Fishburn (1970, Section 3.1) for details, or Debreu (1959, Chapter 4), who discusses conditions for u to be continuous. Debreu uses a "continuity" postulate in place of (2) that, in our setting, requires that if the sequence $[L_i]$ converges (in distribution) to the lottery L_i, and $L_i \prec L_k$, then all but finitely many of the $L_i \prec L_k$. If we extend $\precsim$ to general distributions over R, Debreu's continuity postulate entails countably

(3) The "independence" principle: For all L_i, L_j, and L_k, and for all α $(0 < \alpha \leq 1)$, $L_i \lesssim L_j \iff \alpha L_i + (1-\alpha)L_k \lesssim \alpha L_i + (1-\alpha)L_k$.

Let us examine these postulates for the special case of lotteries on three rewards: $R = (r_1 < r_2 < r_3)$, where the reward r_i is identified with the degenerate lottery having point-mass $P(r_i) = 1$ $(i = 1,2,3)$. Following the excellent presentation by Machina (1982), we arrive at a simple geometric account of what is permitted by expected-utility theory. Figure 20.1 depicts the consequences of postulates (l)–(3).

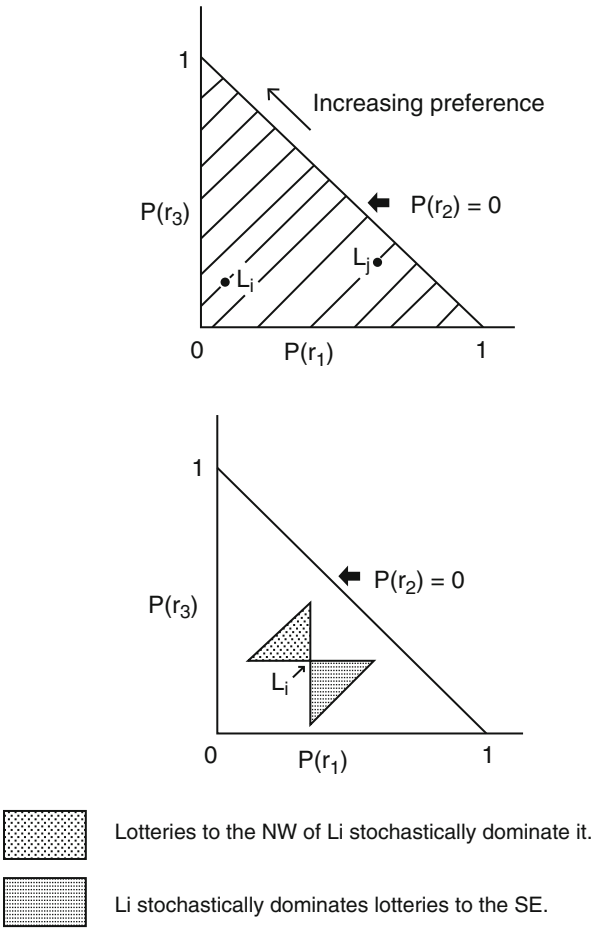


Fig. 20.1 Geometry of cardinal utility with three rewards

additive probability. See Seidenfeld and Schervish (1983) for some discussion of the decision-theoretic features of finitely additive probability.

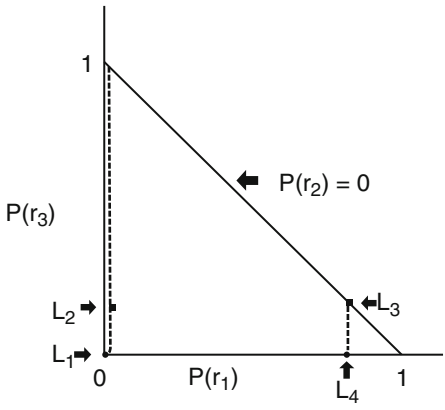


Fig. 20.2 Geometry of the Allais paradox

According to the postulates (l)–(3), indifference curves ($\sim$) over lotteries are parallel, straight lines of (finite) positive slope. L_i is (strictly) preferred to L_j, $L_j < L_i$, is just in case the indifference curve for L_i is to the left of the indifference curve for L_j.

Consider a lottery L_i, as in Fig. 20.2. Stochastic dominance provides a (strict) preference for lotteries to the NW of L_i, whereas L_i is (strictly) preferred to lotteries to its SE.[2] Thus, the indifference lines must have positive slope. Hence, in this setting with lotteries over three rewards, expected-utility theory permits one degree of freedom for preferences, corresponding to the choice of a slope for the lines of indifference.

In a collaborated effort, Seidenfeld et al. (1987, Section 1), we apply this analysis to the selection of "sizes" (α-levels) for statistical tests of a simple null hypothesis against a simple rival hypothesis. The conclusion we derive is the surprising "incoherence" (conflict with expected-utility theory) of the familiar convention to choose statistical tests with a size, e.g., $\alpha = .01$ or $\alpha = .05$, independent of the sample size. This reasoning generalizes that of Lindley (1972, p. 14, where he gives his argument for the special case of "0–1" losses). In a purely "inferential" (nondecision-theoretic) Bayesian treatment for the testing of a simple hypothesis versus a composite alternative, Jeffreys (1971, p. 248) argues for the same caveat about constant α-levels.

[2]Recall, lottery L_2 (first order) stochastically dominates lottery L_1 if L_2 can be obtained from L_1 by shifting probability mass from less to more desirable payoffs. More precisely, L_2 stochastically dominates L_1 if, as a function of increasingly preferred rewards, the cumulative probability distribution for L_2 is everywhere less than (or equal to) the cumulative probability distribution for L_1. Of course, whenever L_2 stochastically dominates L_1, there is a scheme for payoffs, in accord with the two probability measures, where L_2 weakly dominates L_1.

Review of the Allais and Ellsberg "Paradoxes"

3.1

Allais (1953) poses the following question. For the three rewards, $r_1 = \$0$, $r_2 = \$1$ million, and $r_3 = \$5$ million (so $r_1 < r_2 < r_3$), what are your preferences, in the choice between lotteries L_1 and L_2, and in the choice between lotteries L_3 and L_4, where:

$$\begin{aligned} &L_1 - \text{with } P(r_2) = 1 \quad (\$1 \text{ million for certain}), \\ &L_2 - \text{with } P(r_1) = .01, \quad P(r_2) = .89, \quad \text{and } P(r_3) = .10, \\ &L_3 - \text{with } P(r_1) = .90 \text{ and } P(r_3) = .10, \text{ and} \\ &L_4 - \text{with } P(r_1) = .89 \text{ and } P(r_2) = .11? \end{aligned}$$

The common response, choose L_1 over L_2, and L_3 over L_4, violates EU theory (under the assumption that the choices reveal <). This is made evident by an application, Fig. 20.2, of (Machina's) figure 1.

The lines connecting the pairs of lotteries in the two choices are parallel. Thus, regardless of the slope of the parallel, straight-line indifference curves (from Fig. 20.1) imposed on lotteries over the three rewards, either L_2 and L_3 are preferred to their rivals, or else L_1 and L_4 are preferred. EU precludes the common answer to Allais' question.

3.2

Ellsberg's (1961) paradox of preference for lotteries with known risk ($\in M$) over uncertain lotteries ($\notin M$), bearing unknown risk, does not fit the simple mixture-set model, M. We can accommodate Ellsberg-styled problems by generalizing our concept of acts so that an act is a function f from states, a (finite, exhaustive) partition, to distributions on the reward set R. These more general acts are called "horse lotteries" by Anscombe and Aumann (1963). Denote by M' ($\supset M$) the generalized mixture-set for the class of horse lotteries.[3] Then lotteries of known risk belong to this enlarged (mixture) set M' as a special case: they are the "constant" acts. That is, the acts of known risk are those for which f^{-1} is a determinate probability measure.

Let us see how an Ellsberg-styled paradoxical choice violates postulate 3, supposing (1) and (2) obtain, when the postulates are applied to M'. Imagine I

[3]To define the generalized mixture-set M', it suffices to define the operation of convex-combination of two (generalized) lotteries. This is done exactly as in Anscombe and Aumann's (1963) treatment of "horse lotteries." Horse lotteries, the generalized postulates (l)–(3) for horse lotteries, and, with the addition of two minor assumptions (precluding a preference-interaction between payoffs and states), the subjective expected-utility theory that results, are discussed by Fishburn (1970, Chapter 13) and briefly in section Sequential coherence of Levi's decision theory here.

have placed \$10 in one of two pockets, which are otherwise empty. Consider the following three lotteries:

L_{left} – take the contents of my left pocket,
L_{right} – take the contents of my right pocket, and
L_{mix} – take the contents of my left pocket if a "fair" coin lands tails up, and take the contents of my right pocket if the fair coin lands head up.

Lotteries L_{left} and L_{right} are uncertain prospects. Suppose you are indifferent ($\sim$) between these two, which you evaluate as having a sure-dollar equivalent of \$2.50. However, the third option, L_{mix}, is (under the "reduction" postulate) a lottery of known risk. That is, L_{mix} is a lottery with an equal (.5; .5) probability distribution on the two payoffs (\$0, \$10). In the spirit of the Ellsberg paradoxical choice, suppose you evaluate the fair gamble on these two payoffs as having, say, a sure-dollar equivalent of \$4.00. You (strictly) prefer the lottery of known risk, L_{mix}, to either of the two uncertain lotteries. Finally, as the coin flip gives you no relevant information about the location of the \$10, your conditional preferences over the two uncertain lotteries (and their \$2.50 equivalent) are unaffected by the outcome of the coin flip. Then, as L_{mix} is (under reduction) equivalent to the ($\alpha = .5$) convex combination of L_{left} and L_{right}, preference for "risk" over "uncertainty" violates the independence postulate 3, assuming (1) and (2) hold.[4]

In fact, given (1) and (2), this version of the Ellsberg paradox conflicts with a principle (4), (strictly) weaker than principle (3).

(4) Mixture dominance ("betweenness"): Of lotteries L_1 and L_2, if each is (weakly or strictly) preferred (or dispreferred) to a lottery L_3; so, too, each convex combination of L_1 and L_2 is (weakly or strictly) preferred (or dispreferred) to L_3.

[4]These preferences are in conflict with Savage's (1954) "sure-thing" postulate P2. P2 is inconsistent with the following two preferences:

(i) $L_{\text{right}} < L_{\text{mix}}$.
(ii) $L_{\text{left}} \sim L_{\text{right}}$, given the coin lands heads up.

Consider the four-event partition generated by whether the coin lands heads (H) or tails (T), and whether the \$10 is in the left (L) or right (R) pocket. Then, by (i), the first row (below) is preferred to the second. Savage's theory uses "called-off" acts to capture conditional preference. Thus, by (ii), the agent is indifferent between the third and fourth rows.

	HL	HR	TL	TR
L_{mix}	\$10	\$0	\$0	\$10
L_{right}	\$0	\$10	\$0	\$10
$L_{\text{left}} \mid \text{H}$	\$10	\$0	\$0	\$0
$L_{\text{right}} \mid \text{H}$	\$0	\$10	\$0	\$0

In terms of (Machine's) figure 1, mixture dominance entails linear indifference curves. (This follows directly with (4), as then indifference is preserved under convex combinations. In Fig. 20.1, the set of convex combinations of two lotteries graphs as a straight line.) But the conjunction of (1), (2), and (4) does not entail (3). Samuelson's (1950) "Ysidro" ranking, and the "weighted utility" theory of Chew (1981) satisfy (1), (2) and (4) but fail (3).[5] Chew (1983) shows that the Allais paradoxical choices are admitted by his theory. What we find here is that Ellsberg-styled preference for risk over uncertainty cannot be so easily absorbed. In order to admit the Ellsberg-styled paradoxical choices, mixture dominance, (4), too, must fail.

Objections to the Denial of "Independence"

On Failures of "Stochastic Dominance"

Kahneman and Tversky's (1979) intriguing alternative to EU, "Prospect Theory," gives a reconstruction of Allais' paradoxical choice behavior at the expense of the independence postulate. Call a simple lottery *regular* provided not all its payoffs are (strictly) preferred to "status quo." Recall, the ranking of a lottery by expected utility uses the formula:

$$\sum_i P(r_i)\, u(r_i)\,.$$

For regular lotteries, the ranking of a lottery by prospect theory uses the formula:

$$\sum_i \pi\left[P(r_i)\right] \upsilon(r_i)\,,$$

where υ is a value-function for rewards (akin to the utility u), and π is some monotone-increasing function with $\pi(0) = 0$ and $\pi(1) = 1$. Again, let us consider (regular) lotteries on three rewards $r_1 < r_2 < r_3$, where we may take r_1 as status quo. If (and only if) π is linear do we have agreement between prospect theory and EU (for then $\pi(x) = x$ with the scalar constant absorbed into the utility u, defined up to

[5] Let u be a utility on payoffs and assume u is positive. Denote by $E_u(L_1)$ the expected utility of lottery L_1 under utility function u. Denote by ${L_1}^{-1}$ the lottery that has payoffs with (multiplicative) inverse utility to L_1. Samuelson's (1950) "Ysidro" ranking, $\precsim_\gamma$, on lotteries is given by the function

$$Y(L_1) \;=\; \left[E_u(L_1)\,/E_u\left({L_1}^{-1}\right)\right]^5.$$

Not only does $\precsim_\gamma$ satisfy the ordering, Archimedean, and mixture dominance postulates while failing independence but in addition $\precsim_\gamma$ respects stochastic dominance!

positive linear transformations). If π is not linear, so that prospect theory violates independence, stochastic dominance fails too.[6]

I do not know whether this aspect of prospect theory has been subjected to test for its descriptive accuracy. (I find it hard to believe that subjects would prefer a stochastically dominated lottery when the comparison involves just three rewards and the two lotteries involved assign identical probability to the status quo reward r_1.) In any event, it is normatively unacceptable to mandate a violation of stochastic dominance. What, after all, is left of the concern to avoid unnecessary losses (with respect to payoffs) when a theory *requires* a strict preference for an option dominated (on a set of positive probability)?

Thus, we shall examine only those violations of independence that induce preferences consistent with the partial order imposed by stochastic dominance. To that end, with an eye on the anticipated exchange between theories that abandon ordering versus those that abandon independence, I elevate respect for stochastic dominance to the status of a coherence condition.

Definition *A decision rule is* coherent *if* (*i*) *admissible choices under the rule are stochastically undominated, and* (*ii*) *admissibility is preserved under substitution* (*at choice points*) *of "indifferent" options.*

In section "Sequential coherence of Levi's decision theory", I summarize choice-based generalizations of "preference over rewards" (to explicate stochastic dominance) and "indifference over options" without assuming that choice induces a weak-order. However, in terminal decisions, when a choice rule induces an ordering $\lesssim$, condition (ii) adds nothing to (i). (See Sen's (1977) excellent discussion relating properties of choice rules to ordering.) That is, suppose lottery L_2 stochastically dominates L_1. By (i), L_1 is inadmissible when L_2 is available for choice, and $L_1 < L_2$. If, moreover, $L_1 \sim L_3$ and $L_2 \sim L_4$ then $L_3 < L_4$ by properties of $\lesssim$; so that inadmissibility of dominated options is preserved under substitution of indifferents, (ii). Therefore, in nonsequential decisions, and depending upon how indifference is defined without ordering, clause ii only serves as an added restriction on the coherence of choice rules that relax ordering. In sequential decisions the situation is rather different. As shown in section "Sequential incoherence – an example when mixture dominance fails, even though a choice rule induces a weak-order and respects stochastic dominance in nonsequential decisions, it may fail to be sequentially coherent.

The point of clause (ii) is to help identify a standard for evaluating decision rules predicated on the supposition that the agent's values for rewards (and for lotteries over those rewards) are stable over time. That is, this standard of coherence is

[6]The result is elementary and has been noted by many, including Kahneman and Tversky (1979, p. 283–284). Suppose π is not linear so that $\pi(p+q) > \pi(p) + \pi(q)$. Then by letting the value $\upsilon(r_2)$ approach the value $\upsilon(r_1)$, the agent is required (strictly) to prefer L_1: $P_1(r_1) = (1 - [p+q])$, $P_1(r_2) = p + q$, and $P_1(r_1) = 0$ – over L_2: $P_2(r_1) = P_1(r_1)$, $P_2(r_2). = p$, $P_2(r_3) = q$, even though L_2 stochastically dominates L_1. The argument for the other case is similar: $\pi(p+q) < \pi(p) + \pi(q)$.

offered for assessing the performance of a choice rule in sequential decisions when basic values are unchanging. Clause (ii) is not cogent, I would argue, when basic values are subject to revision over time. Then there may be a current preference between two rewards that are (to be) judged indifferent relative to the future, changed values. Thus there is no reason to demand that substitution of "future" indifferents preserves the inadmissibility of what is, by current values, a dominated option.

Of course, the agent's knowledge of events, chance occurrences, and preceding choices inevitably changes in the course of a sequential decision. In fact, these changes in evidence are what makes valuable adaptive experimental designs.

Sequential Incoherence – An Example When Mixture Dominance Fails

Respect for stochastic dominance provides a safeguard that choice over lotteries attends to sure-gains in payoffs. Single stage (nonsequential) decisions are thereby protected from violations of weak dominance over rewards. That is not the case, however, when we attend to sequential decisions. Specifically, coherence in nonsequential decisions, in choices over lotteries, does not entail the sequential version of coherence in choices over plans. This is illustrated by an example.

Consider what happens when mixture dominance (4) fails: *Example*: Let lotteries L_1 and L_2 be indifferent with a sure-dollar equivalent of \$5.00. Suppose, contrary to (4), that an equal ($\alpha = .5$) convex combination of them is strictly preferred with a sure-dollar equivalent of, e.g., \$6.00. Denote this by $L_3 = .5\,L_1 + .5\,L_2 \sim \$6.00$. Then, by continuity of preference for monetary payoffs, there is some fee, $\$\in$, that can be attached to the *payoffs* of L_1 and L_2 (resulting in the lotteries denoted by "$L_1 - \in$" and "$L_2 - \in$") satisfying

$$L_4 = (L_3 - \in) = .5\,(L_1 - \in) + .5\,(L_2 - \in) \sim \$5.75.$$

Also, we can find some dollar prize strictly dispreferred to both of the $\$\in -$ modifications of L_1 and L_2, e.g., suppose

$$\$4.00 < (L_1 - \in)\,, (L_2 - \in)\,.$$

Thus, we have the sequence:

$$\$4.00 < (L_1 - \in)\,, (L_2 - \in) < L_1 \sim L_2 \sim \$5.00 < L_4 \sim \$5.75 < L_3 \sim \$6.00$$

and, by assumption, these preferences respect stochastic dominance in dollar payoffs. So nonsequential choices among these options, according to these preferences, result in no incoherence.

Imagine, however, that an agent with these same preferences faces the following, two-stage sequential decision. Initially (at choice point A), the agent has two (sequential) alternatives, plans 1 and 2. Under sequential plan 1, a fair coin is flipped; (a) if it lands heads up, the agent chooses between L_1 and a dollar prize of \$5.50; and (b) if it lands tails up, the choice is between L_2 and the dollar prize of \$5.50. Under sequential option 2, the fair coin is flipped; (c) if it lands heads up, the agent chooses between $L_1 - \in$ and \$4.00; and (d) if it lands tails up, the choice is between $L_2 - \in$ and \$4.00. (The problem is depicted by Fig. 20.3, where $L_1 \sim L_2 \sim \$5.00 < .5\,L_1 + .5\,L_2 \sim \$6.00$, and where one finds the $\$\in$ fee satisfying: $.5(L_1 - \in) + .5(L_2 - \in) \sim \$5.75$.)

How is the agent to choose between plans 1 and 2? It is clear, I think, that he should face up to what he knows his preferences are at choice nodes B, the choices he faces after the coin is flipped.[7] That is, the agent should assess the two (sequential) plans 1 and 2 in light of what he knows they lead to.

Under (1), if the coin lands heads up (a) he will choose \$5.50 over lottery L_1.[8] And if the coin lands tails up (b) again, he will choose the \$5.50 (over L_2). Thus, from the standpoint of (A), choosing plan 1 leads to a sure payoff of \$5.50.

Under (2), if the coin lands heads up (c) he will choose the lottery $L_1 - \in$ over the dollar reward of \$4.00. And if the coin lands tails up (d), the lottery $L_2 - \in$ is preferred to a sure \$4.00. Hence, from the standpoint of (A), choosing plan 2 leads

[7] Hammond's (1976, Section 3.3) felicitous phrase is that the agent uses "sophisticated" versus "myopic" choice.

[8] McClennen (1986, 1988, forthcoming) sketches a program of "resolute" choice to govern sequential decisions when independence fails. I am not very sure how resolute choice works. Part of my ignorance stems from my inability to find a satisfactory answer to several questions.

As I understand McClennen's notion of resolute choice, the agent's preferences for basic lotteries change across nodes in a sequential decision tree. (Then, the premises of the argument in Section 4 do not obtain.) In terms of the problem depicted in Fig. 4, at node A the agent resolves that he will choose L_1 at (a) of node B, and by so resolving increases its value at (a) of node B above the \$5.50 alternative.

There are several difficulties I find with this proposal. I suspect that the new value of L_1 at (a) will be fixed at \$6.00, and likewise for L_2 at (b). (The details of resolute choice are lacking on this point, but this suspicion is based on the observation that a minor variation in the sequential incoherence argument applies unless these two lotteries change their value from node A to node B as indicated. Just modify the construction so that the rejected cash alternative at B is \$6.00 – δ.) Then the assessed value of \$6.00 for L_3 (a mixture of the lotteries L_1 and L_2, now valued at \$6.00 each) is in accord with postulate (2). Such resolutions mandate that changes in preferences agree, sequentially, with the independence postulate. In terms of *sequential* decisions, is it not the case that resolute choice requires changes in values to agree with the independence postulate?

A second problem with resolute choice directs attention at the reasonableness of these mandatory changes in values. For example, consider the Ellsberg-styled choice problem described in Section 3.2. Cast in a sequential form, under this interpretation of resolute choice, if L_{mix} is most preferred, then the agent is required to increase the value for the option "take the contents of the right pocket," given that the coin lands heads up, over the value it has prior to the coin flip.

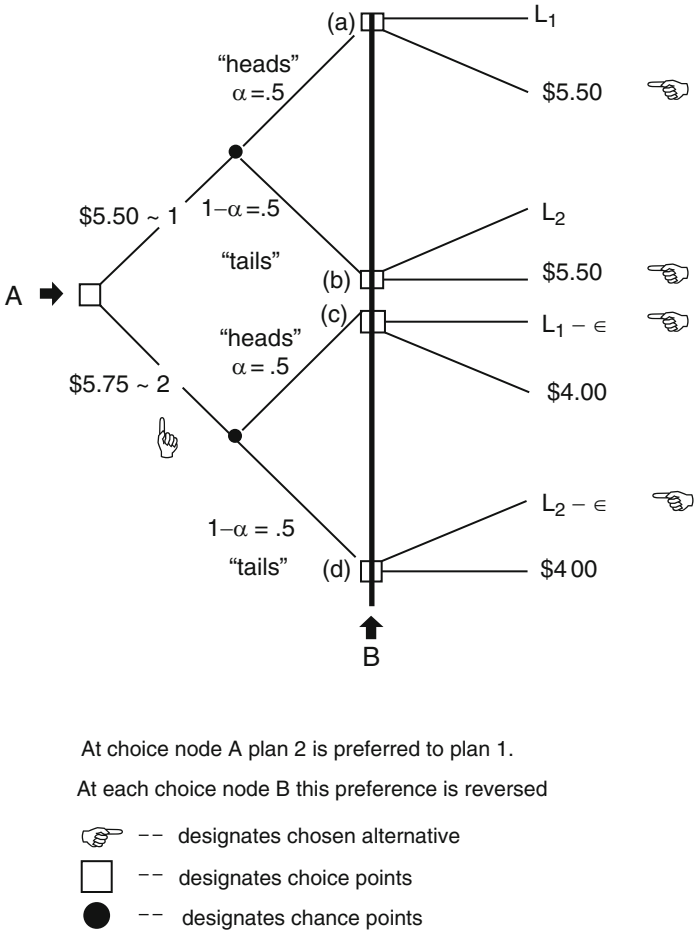


Fig. 20.3 An illustration of sequential incoherence for a failure of mixture of dominance ("betweenness")

to an equal ($\alpha = .5$) convex combination of the two lotteries $L_1 - \in$ and $L_2 - \in$. That is, from the perspective of choice node (A), sequential option 2 yields the lottery L_4, which is valued at \$5.75.

We make this reasoning precise with the following principle.

But the coin flip is irrelevant to a judgment of where the money is. However uncertain the agent is prior to the coin flip, is he not just as uncertain afterwards? Concern with uncertainty in the location of the money is the alleged justification for a failure of independence when comparing the three terminal options: $L_{\text{left}}, L_{\text{right}}, L_{\text{mix}}$, and declaring L_{mix} (strictly) better than the other two. What justifies the preference shift, given the outcome of the coin flip, when L_{right} becomes equivalued with an even-odds lottery over \$10 and \$0 despite the same state of uncertainty about the location of the money before and after the coin flip?

Dynamic Feasibility (DF)
To assess plan p *at a choice node* n_i*, anticipate how you will choose at its* (*potential*) *"future" choice nodes* n_j *and declare infeasible all future alternatives under* p *which are inadmissible at* n_j.

By this account, according to the principle DF, at (*A*) the agent prefers plan 2 over the rival plan 1. At (*A*) plan 2 is worth \$5.75 where plan 1 is worth only \$5.50. However, there is an embarrassment to these preferences. At choice nodes *B*, regardless of the fall of the coin, the agent prefers the choice he makes under plan 1 to what he chooses under plan 2.

If the coin lands heads up, the choice under (1), at (a), \$5.50 is preferred to the choice under (2), at (c), the lottery $L_1 - \in$. Likewise, if the coin lands tails up, the choice under (1), at (b), \$5.50 is preferred to the choice under (2), at (d), $L_2 - \in$. Therefore, though the agent prefers plan 2 to plan 1 initially [at (*A*)], he knows that this preference is reversed at (*B*), regardless of how the coin lands.

Under the indifferences of the ordering postulate and the preferences induced by stochastic dominance, at nodes *B*, the upshot is a contradiction in assessment of the sequential decision problem. The contradiction obtains as follows.

By postulate (1), the agent has a weak ordering of options at choice nodes *B*. There are some sure-dollar equivalents for the choices $L_1 - \in$ and $L_2 - \in$. By simple dominance, the sure-dollar equivalent for $L_1 - \in$ (or for $L_2 - \in$) is less than that for L_1. That is, each of $L_1 - \in$ and $L_2 - \in$ is worth less than \$5.00 and more than \$4.00. Since admissibility at choice nodes (such as nodes *B*) respects indifference, one's anticipation and knowledge [at (*A*)] of one's choices is only up to the level of indifferents. That is, the ordering postulate fixes choices only up to the equivalences of indifference, $\sim$. Thus, substitution of $\sim -$indifferents at choice nodes *B* leaves unchanged choices at those nodes. But plan 2 is dominated by plan 1 when, as dictated by postulate (1), the choices at nodes (c) and (d) [of (*B*)] are switched for their (indifferent) sure-dollar equivalents.

Of course, given the replacements of $L_1 - \in$ and $L_2 - \in$ by their sure-dollar equivalents at (*A*) plan 1 is preferred to plan 2 by dominance. (See Fig. 20.4, where, as in Fig. 20.3, $L_1 \sim L_2 \sim \$5.00 < .5\,L_1 + .5\,L_2 \sim \$6.00$, and where one finds the \$$\in$ fee satisfying: $.5(L_1 - \in) + .5(L_2 - \in) \sim \$5.75$.) Hence, subject to DF, applications of postulate (1) – ordering – at nodes *B* lead to contradictory conclusions with decisions made at node *A*.

Let us call this contradictory assessment an episode of *sequential incoherence*. That is, subject to Dynamic Feasibility, respect for stochastic dominance is not

Third, by what standards can the agent reassess the merits of a resolution made at an earlier time? In other words, how is the agent to determine whether or not to ignore an earlier judgment: the judgment to commit himself to a resolute future choice and thereby to change his future values. Once the future has arrived, why not instead frame the decision with the current choice node as the initial node without altering basic values? Unless this issue is addressed, the question of how to ratify a resolution is left unanswered, and the problem remains of how to make sense of the earlier resolution once the moment of choice is at hand.

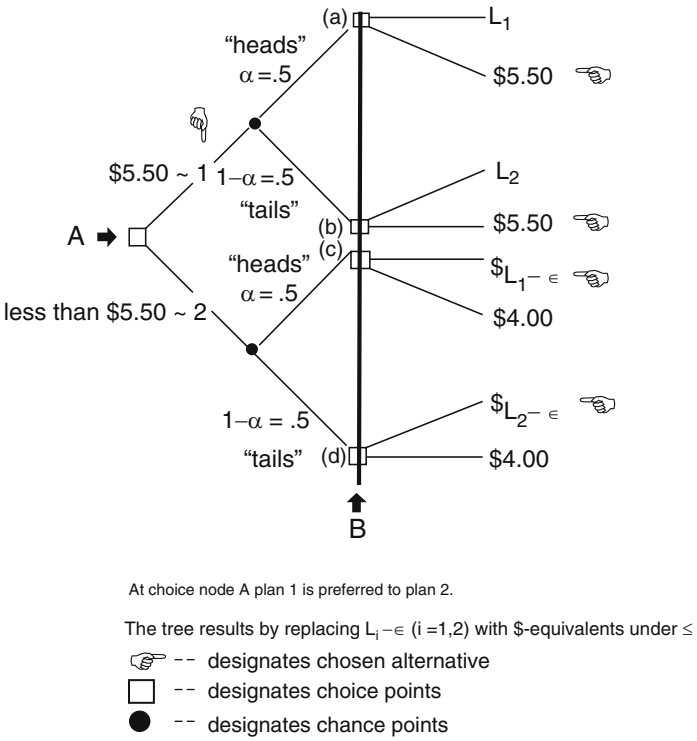


Fig. 20.4 An illustration of sequential incoherence for a failure of mixture of dominance ("betweenness")

preserved under substitution of indifferent alternatives at choice nodes. That is, clause ii of coherence fails with this choice rule.

It is important to understand that, at (*A*) (in the problem depicted in Fig. 20.3), the agent has *no* "terminal" options, no choices of lotteries. In particular, at choice node *A*, the agent does not have the terminal option L_3. Nor does he have any of the terminal options corresponding to the other seven lotteries that arise from an equal ($\alpha = .5$) convex combination of the options available to him at nodes *B*. Thus, at (*A*), he does not have the choice of \$5.50 outright. What he does have as a choice at (*A*) is plan 1, which, by DF, he equates with a certain \$5.50. But plan 1 calls for decisions at nodes *B*, depending upon how the coin lands, and these subsequent choices are not to be ignored at (*A*). The principle of Dynamic Feasibility achieves a limited reduction of plans to terminal options.

In the language of game theory (Luce and Raiffa 1957, chapter 3), the sequential decision problem (above) is in *extensive* form. What we learn from this problem is that, when mixture dominance fails, even with DF, sequential decisions in extensive form are not equivalent to the normal form one-stage (nonsequential) decisions that result by ignoring subsequent choice nodes like (*B*). In a normal form version of this

sequential problem each of the two plans is represented by a set of four lotteries. In normal form, the decision is among the eight lotteries:

$$[L_3, (.5 \cdot \$5.50 + .5L_2), (.5L_1 + .5 \cdot \$5.50), \$5.50, \ldots, \$4.00].$$

Of these L_3 is most preferred, say. In normal form, then, plan 1 is chosen and is valued at \$6.00 ($\sim L_3$). However, the argument offered in this section (establishing sequential incoherence) does not presume the equivalence of extensive and normal forms.[9]

In the sequential problem, at node A, the agent knows L_3 is not available to him under plan 1. This is because he knows that (at nodes B) the dollar prize (\$5.50) is preferred to each of the lotteries L_1 and L_2. Under these preferences, the choice of plan 1 at (A) in the hope that L_1 will be chosen if heads and L_2 if tails is a pipe dream – mere wishful thinking that is brought up short by Dynamic Feasibility.[10]

Sequential Incoherence from Failures of Independence

What is the relation between failures of independence and episodes of sequential incoherence? An answer is given by the central critical result of this essay:

Theorem *If $\lesssim$ is a weak-order (1) of simple lotteries satisfying the Archimedean postulate (2) with sure-dollar equivalents for lotteries, and if $\lesssim$ respects stochastic dominance in payoffs ("a greater chance at more is better"), then a failure of independence, (3), entails an episode of sequential incoherence.*

[9]By contrast, Raiffa's (1968, pp. 83–85) classic objection to the failure of independence in the Allais paradox depends upon a reduction of extensive to normal forms. Also, in his interesting discussion, Hammond (1984) requires the equivalence of extensive and normal forms through his postulate of "consequentialism" in decision trees. These authors defend a strict expected-utility theory, in which the equivalence obtains. Likewise, LaValle and Wapman (1986) argue for the independence postulate with the aid of the assumption that extensive and normal forms are equivalent.

The analysis offered here does not presume this equivalence, nor does avoidance of sequential incoherence entail this equivalence, since, e.g., it is not satisfied in Levi's theory either – though his theory avoids such incoherence. Hence, for the purpose of separating decision theories without independence from those without ordering, it is critical to avoid equating extensive and normal forms of decision problems.

[10]One may propose that, by force of will, an agent can introduce new terminal options at an initial choice node, corresponding to the "normal" form version of a sequential decision given in "extensive" form. Thus, for the problem depicted in Fig. 20.3, the assumption is that the agent can create the terminal option L_3 at node A by opting for plan 1 at A and then choosing L_1 at (a) and L_2 at (b).

Whatever the merit of this proposal, it does not apply to the sequential decisions discussed here, since, by stipulation, the agent cannot avoid reconsideration at nodes B. There may be some problems in which agents can create new terminal options at will, but that is not a luxury we freely enjoy. Sometimes we have desirable terminal options and sometimes we can only plan. (See Levi's [1980, Chapter 17] interesting account of "using data as input" for more on this subject.)

Proof The proof is given in two cases. (The argument for the second case uses the full assumption of stochastic dominance rather than the weaker assumption [used in Case 20.1] that < respects simple dominance in dollar payoffs.)

Case 20.1 Let $L_1 \lesssim L_2$, yet for some L_3 and α, $\alpha L_2 + (1-\alpha)L_3 < \alpha L_1 + (1-\alpha)L_3$. Let $\$X \sim \alpha L_2 + (1-\alpha)L_3$, $\$Z \sim \alpha L_1 + (1-\alpha)L_3$, and $\$\mathcal{U} \sim L_3$, with $X < Z$. By our assumptions of a weak order for preference, continuity in dollar payoffs, and the strict preference for more (over less) money, there is some $2\in$ fee and amount $\$Y$ for which:

$$\$X < \alpha\,(L_1 - 2\in) + (1-\alpha)\,(L_3 - 2\in) \sim \$Y < \$Z.$$

Next, choose fees $\$\gamma$ and $\$\delta$ so that:

$$L_2 - \gamma < L_1 - 2\in \text{ and } \$\,(\mathcal{U} - 2\delta) < L_3 - 2\in .$$

If we consider the sequential decision problem whose "tree" is depicted in Fig. 20.5 for Case 20.1, we discover by the same reasoning we used in the example above:

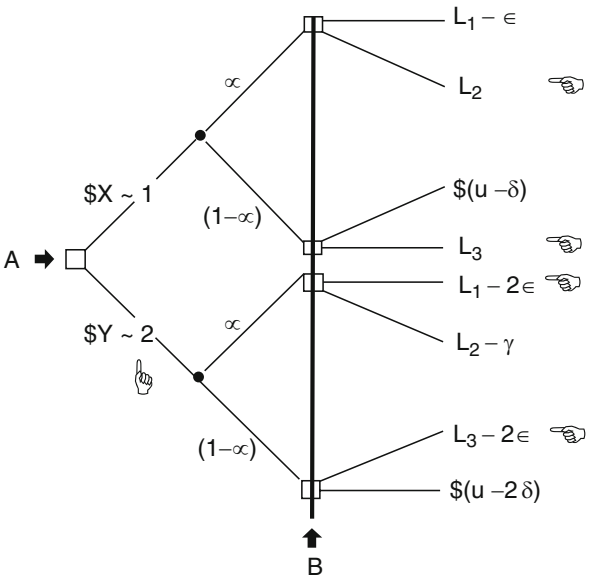


Fig. 20.5 Sequential incoherence for failures of "independence": Case 20.1

At node A, plan 1 is valued at \$$X$, whereas plan 2 is valued at \$$Y$. Thus, at node A, plan 2 is the preferred choice.

But at nodes B, regardless of which "chance event" occurs, the favored option under plan 1 is preferred to the favored option under plan 2. Thus, the preferences leading to a failure of independence in Case 20.1 succumb to sequential incoherence. An application of indifferences (from the ordering postulate 1 at nodes B leads to an inconsistent evaluation, at (A), of the sequential plans 1 and 2.

Case 20.2 $L_1 < L_2$, yet there are L_3 and $\alpha > 0$ with $\alpha L_1 + (1-\alpha)L_3 \sim \alpha L_2 + (1-\alpha)L_3$. Let \$$\mathcal{U} \sim L_3$ and \$$Z \sim \alpha L_1 + (1-\alpha)L_3$. Then choose an \$$\in$ fee so that $L_1 < L_2 - \in$. Let \$$X \sim \alpha(L_2 - \in) + (1-\alpha)L_3$. Then by stochastic dominance, $X < Z$. Next, by continuity, choose a \$$\delta$ fee to satisfy \$$Y \sim \alpha(L_1 - \delta) + (1-\alpha)(L_3 - \delta)$, where $X < Y < Z$. Choose an integer n so that $L_2 - n\in < L_1 - \delta$. Finally, find any \$$\gamma$ fee and choose an integer m so that \$$(u - m\gamma) < L_3 - \delta$.

Consider a sequential decision problem whose "tree" is depicted in Fig. 20.6 for Case 20.2. Once again we find an episode of sequential incoherence as:

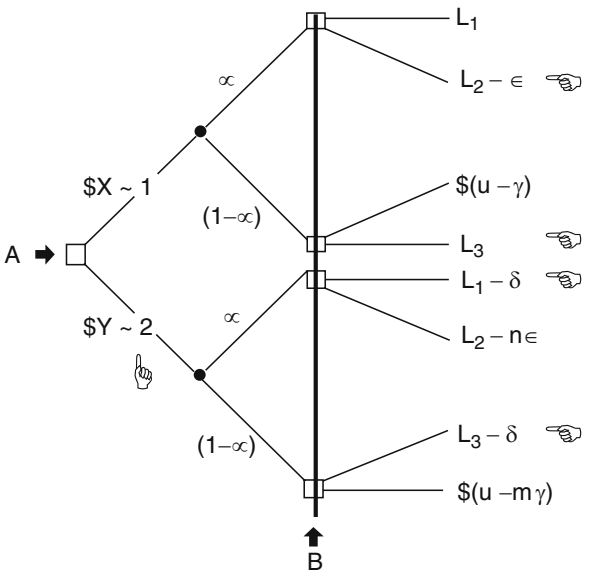


Fig. 20.6 Sequential incoherence for failures of "independence": Case 20.2

At (*A*), plan 1 is valued at $\$X$, whereas plan 2 is valued at $\$Y$. Thus, at node *A* plan 2 is the preferred choice.
At node *B*, regardless of which chance event occurs, the favored option under plan 1 is preferred to the favored option under plan 2. Thus, the preferences leading to a failure of independence in Case 20.2 succumb to sequential incoherence.

Concluding Remark

Can familiar "Dutch Book" arguments (de Finetti 1974; Shimony 1955) be used to duplicate the results obtained here? Do the considerations of book establish sequential incoherence when independence fails? I think they do not.

The book arguments require an assumption that the concatenation (conjunction) of favorable or indifferent or unfavorable gambles is, again, a favorable or indifferent or unfavorable gamble. That is, the book arguments presume payoffs with a simple, additive, utility-like structure. The existence of such commodities does not follow from a (subjective) expected-utility theory, like Savage's. And rivals to EU, such as Samuelson's (1950) "Ysidro" ranking, can fail to satisfy this assumption though they respect stochastic dominance in lotteries. Thus, in light of this assumption about combinations of bets, the Dutch Book argument is not neutral to the dispute over coherence of preference when independence fails. (Of course, that debate is not what Dutch Book arguments are designed for.)[11]

This objection to the use of a book argument does not apply to the analysis presented here. The argument for sequential incoherence is not predicated on the Dutch Book premise about concatenations of favorable gambles. That assumption is replaced by a weaker one, to wit: $\lesssim$ respects stochastic dominance in $-rewards. There is no mystery why the weakening is possible. Here, we avoid the central question addressed by the Dutch Book argument: When are betting odds subjective probabilities? The book arguments pursue the representation of coherent belief as probabilities, given a particular valuation for combinations of payoffs. Instead, the spotlight here is placed on the notion of coherent sequential preference, given a preference (a weak ordering) of lotteries with canonical probabilities.

[11] See Frederic Schick's "Dutch Bookies and Money Pumps" (1986) for discussion of the import of this concatenation assumption in the Dutch Book argument. Its abuse in certain "intertemporal" versions of Dutch Book is discussed in Levi (1987).

Summary

Attempts to generalize EU by denying independence, while retaining the ordering and Archimedean postulates, fail the test of coherence in simple sequential choices over lotteries with dollar rewards.

Sequential Coherence of Levi's Decision Theory

A detailed analysis of Levi's Decision Theory (LDT), a theory without the ordering postulate, is beyond the scope of this essay. Here, instead, I shall merely report the central results that establish coherence of LDT in sequential choices over horse lotteries, a setting where both values (utility/security) and beliefs (probability) may be indeterminate. (I give proofs of these results in a technical report Seidenfeld (1987)).

To begin with, consider the following choice-based generalizations of the concepts: indifference, preference, and stochastic dominance. These generalizations are intended to apply in the domain of horse-lottery options, regardless of whether or not a decision rule induces a weak-ordering of acts.

The notational abbreviations I use are these. An option is denoted by o_i and, since acts are functions from states to outcomes, also by the function on states $o_i(s)$. Sets of feasible options are denoted by O, and the admissible options (according to a decision rule) from a feasible set O are denoted by the function C[O].

Call two options $\approx$ –*indifferent*, if and only if, whenever both are available either both are admissible or neither is.

Definition

$$o_1 \approx o_2 \iff \forall\,(O)\ (\{o_1, o_2\} \subset O \Rightarrow (\,o_1 \in C\,[O] \iff o_2 \in C\,[O])).$$

When a choice rule induces a weak-order, denoted by $\lesssim$, then $\approx$ is just the symmetrized $\sim$ relation: $(o_1 \sim o_2) \iff (o_1 \lesssim o_2)$ *and* $(o_2 \lesssim o_1)$.

Next, define a choice-based relation of categorical preference over rewards using a restricted version of Anscombe and Aumann's (1963, p. 201) "Assumption 1," modified to avoid the ordering postulate. [This assumption is part of the value-neutrality of states. See Drèze's (1985) monograph for a helpful discussion of related issues.] Let o_1 and o_2 be two horse lotteries that differ only in that o_2 awards reward r_2 on states where o_1 awards reward r_1. Then reward r_2 is categorically preferred to reward r_1 just in case o_1 is inadmissible whenever o_2 is available. In symbols,

Definition *Reward* r_2 *is* categorically preferred *to reward* $r_1 \iff \forall(O)\ \forall(o_1 \neq o_2)\ (\forall(s)[o_1(s) = o_2(s) \ \vee\ (o_1(s) = r_1\ \&\ o_2(s) = r_2)]\ \&\ o_2 \in O \Rightarrow o_1 \notin C[O])$.

Third, we need to make precise a suitable generalization of stochastic dominance among lotteries. Recall, when there is an ordering ($\lesssim$) of rewards, lottery L_2 stochastically dominates lottery L_1 if L_2 can be obtained from L_1 by shifting some distribution mass from less to more preferred rewards. Thus, L_2 stochastically dominates L_1 just in case $\alpha L_2 + (1-\alpha)L_3$ stochastically dominates $\alpha L_1 + (1-\alpha)L_3$ $(0 < \alpha \leq 1)$. Rely on this biconditional to formulate the following $\prec$*−dominance* relation over horse lotteries, defined for an arbitrary choice function C.

Definition $o_1 \prec o_2 \iff \forall(O)\forall(o)\forall(\alpha > 0)\ ((\alpha o_2 + (1-\alpha)o) \in O \Rightarrow (\alpha o_1 + (1-\alpha)o) \notin C[O])$. Trivially, if $o_2 \prec$*−dominates* o_1, *then* (*let* $\alpha = 1$) o_1 *is inadmissible whenever* o_2 *is available.*
Sequential coherence for a decision rule requires, then, that

(i) shifting distributional mass from categorically less to categorically more preferred rewards produces an $\prec$−dominating option, and
(ii) the inadmissibility of $\prec$−dominated options is preserved under substitution (at choice nodes) of $\approx$−indifferents.

To see why LDT is sequentially coherent, recall that admissibility in Levi's (1974, 1980) decision theory is determined by a two-tiered lexicographic rule. The first tier is "*E*-admissibility." An option is *E*-admissible in the set O of feasible options, provided it maximizes expected utility (over O) for some pair $(P, \mathcal{U})$ – where P is a personal probability in the (convex) set **P** that represents the agent's beliefs about the states, and where $\mathcal{U}$ is a (cardinal) utility in the (convex) set **U** that represents one aspect of the agent's values over the rewards.

Definition o *is* E *-admissible* $\iff \exists(P,U)\ \forall(o' \in O)\ E_{PU}[o] \geq E_{PU}[o']$.[12]

The second tier in admissibility requires that an option be "S-admissible." This condition demands of an option that it be *E*-admissible and maximize a "security" index among the *E*-admissible options. The security of an option reflects yet another aspect of the agent's value structure. For purposes of this section, the notion of security is illustrated with three (real-valued) varieties:

$\mathrm{Sec}_o[o] = O$ – a vacuous standard, where all options have equal security;
$\mathrm{sec}_1[o] = \text{infemum}_{U,\,R_o}\, \mathcal{U}[r]$, where R_o is the set of possible outcomes (rewards) for option o and $r \in R_o$. Thus, sec_1 is security indexed by the "worst" possible reward, a maximin consideration;

[12]Levi (1980, Section 5.6) offers a novel rule, here called rule', for determining expectation-inequalities when the partition of states, π, is finite but when events may have subjective probability 0. The motive for this emendation is to extend the applicability of "called-off" bets (Shimony 1955) to include a definition of conditional probability given an event of (unconditional) probability 0. Also, it extends the range of cases where a weak-dominance relation determines a strict preference.

Given a probability/utility pair $(P, \mathcal{U})$, maximizing †-expected utility (with rule') includes a weak-order that satisfies the independence axiom, though, †-expectations may fail to admit a real-valued representation, i.e., the "Archimedean" axiom is not then valid. Under rule†, given a pair $(P, \mathcal{U})$, †-expectations are represented by a lexicographic ordering of a vector-valued quantity.

$\text{sec}_2[o] = \text{infemum}_{P\times U}\, E_{P,U}[o]$ – security indexed by the least expected utility, also called the "Γ-minimax" level for option o. (I avoid the minor details of defining sec_2 when events have probability O and expectations are determined by rule†, as reported in note 12.)

Thus, an option is admissible in LDT just in case it is both E-admissible and maximizes security among those that likewise are E-admissible. As a special case, when security is vacuous or is indexed by sec_2, and when both **P** and **U** are unit sets, C[O] is the subset of options that maximize subjective expected utility: the strict Bayesian framework. Then admissibility satisfies the ordering and independence postulates (and the Archimedean axiom, too, provided events have positive probability or rule† is not used).

The next few results, stated without proof, provide the core of the argument for sequential coherence of LDT. Condition i of coherence in nonsequential decisions, and more, is shown by the following.

Theorem 20.1 *If* o_2 *can be obtained from* o_1 *by shifting distribution masses from categorically less to more preferred rewards, then* $o_1 < o_2$, *and thus* o_1 *is inadmissible whenever* o_2 *is available*.

The theorem follows directly from a useful lemma about < and categorical preference in LDT.

Lemma $o_1 \prec o_2 \iff \forall(P, U) E_{PU}(o_1) < E_{PU}(o_2)$. *Thus,* r_2 *is categorically preferred to* $r_1 \iff \forall(U)\, U(r_1) < U(r_2)$.

Thus, both $\prec$-dominance and categorical preference are strict partial orders, being irreflexive and transitive. In addition, following obviously from its definition, $\prec$-dominance satisfies the independence axiom. [Note the term "categorical preference" is borrowed from Levi (1986b, p. 91). The lemma provides the collateral for this conceptual loan.]

Condition ii of coherence for LDT in nonsequential decisions is shown by a result that two options are $\approx$ −related exactly when they have the same security index and are indiscernible by expectations:

Theorem 20.2 $o_1 \approx o_2 \iff (sec[o_1] = sec[o_2]\ \&\ \forall(P,U)(E_{PU}[o_1] = E_{PU}[o_2])$. *Thus,* $\approx$ *is an equivalence relation and, in nonsequential decisions, admissibility is preserved under substitution of* $\approx$ *−indifferent options*.

In order to extend the analysis to sequential decisions, the notion of $\approx$ − indifference is generalized to include conditional assessments, conditional upon the occurrence of either "chance" or "event" outcomes. The next corollary is elementary and indicates how conditional $\approx$ −indifference applies when for instance, choice nodes follow chance nodes.

Corollary 20.1 $\approx$ *is preserved under chance mixtures*,

$$\forall\,(i, j, k \text{ and } \alpha)\, o_i \approx o_i \Rightarrow (\alpha o_i + (1-\alpha)\, o_k \approx \alpha o_i + (1-\alpha)\, o_k)\,.$$

Under the principle of Dynamic Feasibility (which provides a limited reduction of sequential plans to nonsequential options), these findings combine to support the desired conclusion.

Sequential coherence for LDT

(i) Admissible options are $\prec$ −undominated among the dynamically feasible alternatives, and
(ii) Provided the agent updates his beliefs by Bayes' rule and does not change his value structure, admissibility is preserved under the substitution of $\approx$ −indifferent alternatives at choice nodes.

Conclusions About Relaxing Independence or Ordering

I have argued that decision theories that relax only the independence postulate succumb to sequential incoherence. That is, such programs face the embarrassment of choosing stochastically dominated options when, in simple two-stage sequential decisions, dollar equivalents are substituted for their indifferent options at terminal choice nodes. Moreover, the criticism does *not* presume an equivalence between sequential decisions (in extensive form) and their normal form reductions; instead, all decisions are subject to a principle of Dynamic Feasibility.

In section "Sequential coherence of Levi's decision theory", I generalize sequential coherence to choice rules that may not induce an ordering of options by preference. Also, I outline reasons for the claim that Levi's Decision Theory is sequentially coherent (in a setting where both belief and value are subject to indeterminacy). Since Levi's theory is one that fails the ordering postulate, the combined results establish a demarcation between these two strategies for relaxing traditional (subjective) expected-utility theory. The difference is that only one of the two approaches is sequentially coherent.

Acknowledgments I have benefitted from discussions with each of the following about the problems addressed in this essay: W. Harper. J. Kadane, M. Machina, P. Maher, E. F. McClennen, M. Schervish; and I am especially grateful to I. Levi.

Discussions

Editors' Note

Subjective expected-utility theory provides simple and powerful guidance concerning how to make rational decisions in circumstances involving risk. Yet actual decision making often fails, as has been well known for decades, to conform to the theory's recommendations. If subjective expected-utility theory represents the

ideal of rational behavior, these failures may simply show that people often behave irrationally. Yet if the gap between ideal and actual behavior is too wide, or if behavior that on the best analysis we can make is rational but not consistent with subjective expected-utility theory, then we may come to doubt some of the axioms of the theory. Two main lines of revision have been suggested: either weakening the "ordering" axiom that requires preferences to be complete or surrendering the so-called independence principle. Although the issues are highly abstract and somewhat technical, the stakes are high; subjective expected-utility theory is critical to contemporary economic thought concerning rational conduct in public as well as private affairs.

In the preceding article, "Decision Theory without 'Independence' or without 'Ordering': What Is the Difference?" Teddy Seidenfeld argued for the sacrifice of ordering rather than independence by attempting to show that abandoning the latter leads to a kind of sequential incoherence in decision making that will not result from one specific proposal (Isaac Levi's) for abandoning ordering. In their comments in this section, Edward McClennen, who supports surrendering the independence postulate rather than ordering, and Peter Hammond, who argues against any weakening of subjective expected-utility theory, discuss Seidenfeld's argument from their quite different theoretical perspectives.

Economics and Philosophy, 4, 1988, 292–297. Printed in the United States of America.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école americaine. *Econometrica, 21*, 503–546.

Allais, M. (1979). The so-called Allais Paradox and rational decisions under uncertainty. In M. Allais & O. Hagen (Eds.), *Expected utility hypotheses and the Allais Paradox* (pp. 437–681). Dordrecht: Reidel.

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of Mathematical Statistics, 34*, 199–205.

Bell, D. (1982). Regret in decision making under uncertainty. *Operations Research, 20*, 961–981.

Bell, D., & Raiffa, H. (1979). *Decision regret: A component of risk aversion*. Unpublished manuscript, Harvard University.

Chew, S. H. (1981). *A mixture set axiomatization of weighted utility theory* (4th revision). Tuscon: Department of Economics, University of Arizona.

Chew, S. H. (1983). *A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais Paradox.* Tuscon: Department of Economics, University of Arizona.

Chew, S. H., & MacCrimmon, K. R. (1979). *Alpha-Nu choice theory: A generalization of expected utility theory* (University of British Columbia Working Paper).

de Finetti, B. (1974). *Theory of probability* (2 vols.). New York: Wiley.

Debreu, G. (1959). *Theory of value*. New Haven: Yale University Press.

Drèze, J. H. (1985). *Decision theory with moral hazard and state-dependent preferences* (CORE discussion paper #8545). Belgium: Center for Operations Research and Econometrics, Université Catholique de Louvain.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics, 75*, 643–699.
Fishburn, P. C. (1970). *Utility theory for decision making*. New York: Kriefer Publishing Company.
Fishburn, P. C. (1981). An axiomatic characterization of skew-symmetric bilinear functionals, with applications to utility theory. *Economic Letters, 8*, 311–313.
Fishburn, P. C. (1983). Nontransitive measurable utility. *Journal of Mathematical Psychology, 26*, 31–67.
Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society B, 14*, 107–114.
Hammond, P. J. (1976). Changing tastes and coherent dynamic choice. *Review of Economic Studies, 43*, 159–173.
Hammond, P. J. (1984). *Consequentialist behaviour in decision trees is Bayesian rational*. Stanford University.
Herstein, I. N., & Milnor, J. (1953). An axiomatic approach to measurable utility. *Econometrica, 21*, 291–297.
Jeffreys, H. (1971). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.
Jensen, N. E. (1967). An introduction to Bernoullian utility theory: I. Utility functions. *Swedish Journal of Economics, 69*, 163–183.
Kadane, J. (1986). Toward a more ethical clinical trial. *Journal of Medicine and Philosophy, 11*, 385–404.
Kadane, J., & Sedransk, N. (1980). Toward a more ethical clinical trial. In J. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 329–338). Valencia: University Press.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.
LaValle, I. H., & Wapman, K. R. (1986). Rolling back decision trees requires the independence axiom? *Management Science, 32*(3), 382–385.
Levi, I. (1974). On indeterminate probabilities. *Journal of Philosophy, 71*, 391–418.
Levi, I. (1980). *The enterprise of knowledge*. Cambridge: MIT Press.
Levi, I. (1986a). The paradoxes of Allais and Ellsberg. *Economics and Philosophy, 2*, 23–53.
Levi, I. (1986b). *Hard choices: Decision making under unresolved conflict*. Cambridge: Cambridge University Press.
Levi, I. (1987). The demons of decision. *The Monist, 70*, 193–211.
Lindley, D. V. (1972). *Bayesian statistics: A review*. Philadelphia: SIAM.
Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal, 92*, 805–824.
Luce, D., & Raiffa, H. (1957). *Games and decisions*. New York: Wiley.
Machina, M. (1982). Expected utility analysis without the independence axiom. *Econometrica, 50*, 277–323.
Machina, M. (1983, October). *The economic theory of individual behavior toward risk: Theory, evidence and new directions* (Technical report #433). San Diego: Department of Economics, University of California.
McClennen, E. F. (1983). Sure-thing doubts. In B. Stigum & F. Wenstop (Eds.), *Foundations of utility and risk theory with applications* (pp. 117–136). Dordrecht: Reidel.
McClennen, E. F. (1986). Prisoner's dilemma and resolute choice. In R. Campbell & L. Sowden (Eds.), *Paradoxes of rationality and cooperation* (pp. 94–104). Vancouver: University of British Columbia Press.
McClennen, E. F. (1988). Dynamic choice and rationality. In B. R. Munier (Ed.), *Risk, decision, and rationality* (pp. 517–536). Dordrecht: Reidel.
McClennen, E. F. (forthcoming). *Rationality and dynamic choice: Foundational explorations*. Cambridge: Cambridge University Press.
Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainly*. Reading: Addison-Wesley.
Samuelson, P. (1950). Probability and the attempts to measure utility. *Economic Review, 1*, 167–173.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
Schick, F. (1984). *Having reasons*. Princeton: Princeton University Press.
Schick, F. (1986). Dutch bookies and money pumps. *Journal of Philosophy, 83*, 112–119.
Seidenfeld, T. (1987). *Sequential coherence without the 'ordering' postulate: Levi's decision theory* (Technical report). Pittsburgh: Department of Philosophy, Carnegie-Mellon University.
Seidenfeld, T., & Schervish, M. (1983). A conflict between finite additivity and avoiding Dutch book. *Philosophy of Science, 50*, 398–412.
Seidenfeld, T., Schervish, M., & Kadane, J. (1987). Decisions without ordering. In W. Sieg (Ed.), *Acting and reflecting*. Dordrecht: Reidel (Forthcoming).
Sen, A. K. (1977). Social choice theory: A re-examination. *Econometrica, 45*, 53–89.
Shimony, A. (1955). Coherence and the axioms of confirmation. *Journal of Symbolic Logic, 20*, 1–28.
Smith, C. A. B. (1961). Consistency in statistical inference and decision. *Journal of the Royal Statistical Society B, 23*, 1–25.
Suppes, P. (1974). The measurement of belief. *Journal of the Royal Statistical Society B, 36*, 160–175.
von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton: Princeton University Press.
Walley, P., & Fine, T. L. (1979). Varieties of modal (classificatory) and comparative probability. *Synthese, 41*, 321–374.
Wolfenson, M., & Fine, T. (1982). Bayes-like decision making with upper and lower probabilities. *Journal of the American Statistical Association, 77*, 80–88.

Chapter 21
Ambiguity and the Bayesian Paradigm

Itzhak Gilboa and Massimo Marinacci

Introduction

Varying Probability Estimates

John and Lisa are offered additional insurance against the risk of a heart disease. They would like to know the probability of developing such a disease over the next 10 years. The happy couple shares some key medical parameters: they are 70 years old, smoke, and never had a blood pressure problem. A few tests show that both have a total cholesterol level of 310 mg/dL, with HDL-C (good cholesterol) of 45 mg/dL, and that their systolic blood pressure is 130. Googling "heart disease risk calculator", they find several sites that allow them to calculate their risk. The results (May 2010) are:

	John (%)	Lisa (%)
Mayo Clinic	25	11
National Cholesterol Education Program	27	21
American Heart Association	25	11
Medical College of Wisconsin	53	27
University of Maryland Heart Center	50	27

I. Gilboa (✉)
HEC, Paris, France

Tel-Aviv University, Tel Aviv, Israel
e-mail: tzachigilboa@gmail.com

M. Marinacci
Università Bocconi, Milano, Italy
e-mail: massimo.marinacci@unibocconi.it

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_21

As we can see from the table, the estimates vary substantially: the highest for John is 100 % higher than the lowest, whereas for Lisa the ratio is 5:2. Opinion diverge in these examples, even though there are based on many causally independent observations that allow the use of statistical techniques such as logistic regression. However, in many important economic questions, such as the extent of global warming, there are very few past events to rely on. Further, many events, such as revolutions and financial crises, cannot be assumed independent of past observations. Thus, it appears that for many events of interest one cannot define an objective, agreed-upon probability.

Does Rationality Necessitate Probability?

Since the mid-twentieth century, economic theory is dominated by the Bayesian paradigm, which holds that any source of uncertainty can and should be quantified probabilistically.[1] According to this view, John and Lisa should have well-defined probabilities that they will develop a heart disease within the next 10 years, as should Mary for the temperature distribution anywhere on the globe 5 years hence. But where should John, Lisa, or Mary get these probabilities from? If they are to consult experts, they will typically obtain different estimates. Which experts are they to believe? Should they compute an average of the experts' estimates, and, if so, how much weight should each expert have in this average?

The standard line of reasoning of the Bayesian approach is that, in the absence of objective probabilities, the decision maker (DM, for short) should have her own, *subjective* probabilities, and that these probabilities should guide her decisions. Moreover, the remarkable axiomatic derivations of the Bayesian approach (culminating in Savage (1954)), show that axioms that appear very compelling necessitate that the DM behave as if she maximized expected utility relative to a certain probability measure, which is interpreted as her subjective probability. Thus, the axiomatic foundations basically say, "Even if you don't know what the probabilities are, you should better adopt some probabilities and make decisions in accordance with them, as this is the only way to satisfy the axioms."

There is a heated debate regarding the claim that rationality necessitates Bayesian beliefs. Knight (1921) and Keynes (1921, 1937) argued that not all sources of uncertainty can be probabilistically quantified. Knight suggested to distinguish between "risk", referring to situations described by known or calculable probabilities, and "uncertainty", where probabilities are neither given nor computable. Keynes (1937) wrote,

> By 'uncertain' knowledge, let me explain, I do not mean merely to distinguish what is known for certain from what is only probable. The game of roulette is not subject, in this

[1] As Cyert and DeGroot (1974) write on p. 524 "To the Bayesian, all uncertainty can be represented by probability distributions."

> sense, to uncertainty . . . The sense in which I am using the term is that in which the prospect of a European war is uncertain, or the price of copper and the rate of interest twenty years hence . . . About these matters there is no scientific basis on which to form any calculable probability whatever. We simply do not know.

Gilboa et al. (2008, 2009, 2012) argue that the axiomatic foundations of the Bayesian approach are not as compelling as they seem, and that it may be irrational to follow this approach. In a nutshell, their argument is that the Bayesian approach is limited because of its inability to express ignorance: it requires that the agent express beliefs whenever asked, without being allowed to say "I don't know" . Such an agent may provide arbitrary answers, which are likely to violate the axioms, or adopt a single probability and provide answers based on it. But such a choice would be arbitrary, and therefore a poor candidate for a rational mode of behavior.

Axiomatic derivations such as Savage's may convince the DM that she ought to have a probability, but they do not tell her which probability it makes sense to adopt. If there are no additional guiding principles, an agent who picks a probability measure arbitrarily should ask herself, is it so rational to make weighty decisions based on my arbitrarily-chosen beliefs? If there are good reasons to support my beliefs, others should agree with me, and then the probabilities would be objective. If, however, the probabilities are subjective, and others have different probabilities, what makes me so committed to mine? Wouldn't it be more rational to admit that these beliefs were arbitrarily chosen, and that, in fact, I don't know the probabilities in question?

Outline

The rest of this paper is organized as follows. Section "History and Background" discusses the history and background of the Bayesian approach. It highlights the fact that this approach has probably never been adopted with such religious zeal as it has within economic theory over the past 60 years. Section "Alternative Models" describes several alternatives to the standard Bayesian model. It surveys only a few of these, attempting to show that much of the foundations and machinery of the standard model need not be discarded in order to deal with uncertainty. Section "Ambiguity Aversion" surveys the notion of ambiguity aversion. The updating of non-Bayesian beliefs is discussed in section "Updating Beliefs". Section "Applications" briefly describes some applications of non-Bayesian models. The applications mentioned here are but a few examples of a growing literature. They serve to illustrate how non-Bayesian models may lead to different qualitative predictions than Bayesian ones. A few general comments are provided in section "Conclusion".

History and Background

Early Pioneers

Decision theory was born as a twin brother of probability theory through the works of a few scholars in the sixteenth and seventeenth century, originally motivated by the study of games of chance. Among them the works of Christiaan Huygens (1629–1695) and Blaise Pascal (1623–1662) are particularly relevant. We begin with Pascal, whose footsteps Huygens followed.

Pascal (1670) Since its very early days, probability had two different interpretations: first, it captures the notion of *chance*, referring to relative frequencies of occurrences in experiments that are repeated under the same conditions. This includes the various games of chance that provided the motivation for the early development of the theory. Second, probability can capture the notion of *degree of belief*, even when no randomness is assumed, and when nothing remotely similar to a repeated experiment can be imagined.

It is this second interpretation that, over time, evolved into the Bayesian approach in both decision and probability theory. In this regard, Pascal is perhaps the most important pioneer of probability theory. Though he made early key contributions to the probabilistic modeling of games of chance, it is his famous wager that is mostly relevant here. Roughly at the same time that Descartes and Leibniz were attempting to prove that God existed, Pascal changed the question from the proof of existence to the argument that it is worthwhile to believe in God, an option that he identified with the choice of a pious form of life based on the precepts of the Christian religion.[2] In so doing, he applied the mathematical machinery developed for objective probabilities in games of chance to the subjective question of God's existence, where no repeated experiment interpretation is possible. This led him to informally introduce several major ideas of modern decision theory, including the decision matrix, the notion of dominant strategies, subjective probability, expected utility maximization, and non-unique probability.[3]

Thus, the subjective interpretation of probabilities and their application as a tool to quantify beliefs showed up on the scene more or less as soon as did the objective interpretation and the application to games of chance. Further, as soon as the notion of subjective probability came on stage, it was accompanied by the possibility that this probability might not be known (see Shafer (1986), for related remarks on Bernoulli (1713), who introduced the law of large numbers).

[2] According to Pascal a pious life would ultimately induce faith. Importantly, Pascal did not assume that one can simply choose one's beliefs.

[3] Pascal did not finish his *Pensées*, which appeared in print in 1670, 8 years after his death. The text that was left is notoriously hard to read since he only sketches his thoughts (here we use the 1910 English edition of W. F. Trotter). Our rendering of his argument crucially relies on Hacking (1975)'s interpretation (see Hacking 1975, pp. 63–72; Gilboa 2009, pp. 38–40).

Huygens (1657) In the wake of the early probability discussions of Fermat and Pascal, Huygens (1657) first clearly proposed expected values to evaluate games of fortune.[4] Unlike Pascal's grand theological stand, Huygens only dealt with games of fortune ("cards, dice, wagers, lotteries, etc." as reported in the 1714 English version). Nevertheless, he was well aware of the intellectual depth of his subject.

Huygens' arguments are a bit obscure (at least for the modern reader; see Daston (1995)). His essay has, however, a few remarkable features from our perspective. First, he does not present the expected value criterion as an axiom; rather, he justifies its relevance by starting from more basic principles. For this reason his essay is articulated in a sequence of mathematical propositions that establish the expected value criterion for more and more complicated games. Huygens's propositions can be thus viewed as the very first decision-theoretic representation theorems, in which the relevance of a decision criterion is not viewed as self-evident, but needs to be justified through logical arguments based on first principles.

A second remarkable feature of Huygens' essay is the basic principle, his "postulat" in the English version, which he based his analysis upon. We may call it the principle of equivalent games, in which he assumes that the values of games of chances should be derived through the value of equivalent fair games. Ramsey's assumption of the existence of bets with equally likely outcomes (that he calls "ethically neutral") is an instance of this principle, as well as de Finetti's assumption of the existence of partitions of equally likely events. More recently, the central role that certainty equivalents play in many axiomatic derivations can be viewed as a later instance of Huygens' comparative principle of studying uncertain alternatives by means of benchmark alternatives with suitably simple structures.

Hacking (1975) makes some further observations on the relevance of Huygens' book for the history of subjective probability. We refer the interested reader to his book, with a warning on the difficulty of interpreting some of Huygens' arguments.

Subjective Probabilities and the Axiomatic Approach

Modern decision theory, and in particular the way it models uncertainty, is the result of the pioneering contributions of a truly impressive array of scholars. Some of the finest minds of the first half of last century contributed to the formal modeling of human behavior. Among them, especially remarkable are the works of Frank Plumpton Ramsey (1903–1930) with his early insights on the relations between utilities and subjective probabilities, John von Neumann (1901–1957) and Oskar Morgenstern (1902–1977) with their classic axiomatization of expected utility presented in the 1947 edition of their famous game theory book, Bruno de Finetti (1906–1985) with his seminal contributions to subjective probability, and Leonard J. Savage (1917–1971), who – in an unparalleled conceptual and mathematical tour de

[4] See Ore (1960).

force – integrated von Neumann-Morgenstern's derivation of expected utility with de Finetti's subjective probability.

For our purposes the contributions of de Finetti, Ramsey, and Savage are especially relevant since they shaped modern Bayesian thought and, through it, the modeling of uncertainty in economics. Next we briefly review their landmark contributions.

Ramsey (1926a) A main motivation of Ramsey (1926a) was Keynes (1921)'s logical approach to probability theory, in which the degrees of beliefs in different proposition were connected by necessary/objective relations, called probability relations. Skeptical regarding the existence of such relations, Ramsey argued that degrees of belief should be viewed and studied as subjective entities. To this end, he promoted the behavioral definition of subjective probability as willingness to bet, and claimed that if subjective probabilities, so defined, do not follow standard probability calculus, the individual will make incoherent decisions. These are two central ideas in the methodology of decision theory, which in about the same years were also advocated by Bruno de Finetti.

Specifically, the first tenet of Ramsey's approach is that the only sensible way to measure degrees of beliefs is not through introspection, but by considering them as a basis of action. The second main tenet is that the rules of standard probability calculus correspond to consistent betting behavior. By consistent betting behavior he meant behavior that was not subject to so-called "Dutch books" . Both ideas also appear in de Finetti (1931), and they are in line with the preaching of the logical positivist, culminating in the Received View, first stated by Rudolf Carnap in the 1920s (see Carnap 1923; Suppe 1977). de Finetti explicitly adopted the doctrine of Operationalism (see, e.g., the last chapter of his 1937 article), and saw the elicitation of subjective probabilities through betting behavior as methodologically akin to Vilfredo Pareto's ordinal utility theory based on the elicitation of indifference curves rather than on some psychological entities that could not measured (when data are assumed to be only a weak order over alternatives). Ramsey was motivated by similar methodological concerns, in a Pragmatist perspective,[5] and viewed this approach as akin to what was done in the physical sciences (see, section "Alternative Models" of his article).

de Finetti (1931) Bruno de Finetti, one of the greatest probabilists of the twentieth century, was a key figure in the development of the Bayesian approach. To the best of our knowledge, he was the first to promote the Bayesian approach as an all-encompassing method of reasoning about uncertainty, and he did so with a religious zeal. (See Cifarelli and Regazzini 1996). His two main papers in this regard are probably de Finetti (1931, 1937). In both papers he forcefully emphasized the two key ideas on subjective probabilities that we just discussed in relation with Ramsey's work.

[5]Operationalism started with Bridgman (1927), after Ramsey's articles of 1926a and 1926b.

In his 1931 article in Fundamenta Mathematicae (pp. 320–324), de Finetti first introduced the notion of a *qualitative probability*, which is a binary relation over events interpreted as "at least as likely as" . He viewed this relation as a primitive and did not explicitly relate it to betting behavior. (The connection was made explicit later on by Savage (1954).)

The novelty of de Finetti (1931) was both methodological and scientific. Methodologically, it is one of very first articles that adopted the axiomatic method based on a binary relation $\succsim$ and its numerical representation derived from suitable axioms on $\succsim$.[6] Scientifically, he provided the first result that axiomatized subjective probability, thereby establishing one of the two pillars which Savage's great synthesis relied upon.[7]

Savage (1954) de Finetti's derivation of subjective probability was conceptually complementary with von Neumann and Morgenstern's (vNM 1947) derivation of expected utility maximization under risk, which assumed known numerical probability measures. The integration of de Finetti's subjective probability with vNM's expected utility was achieved by Savage's (1954) book, which derived subjective expected utility maximization when neither probabilities nor utilities were given. For a description and interpretation of Savage's result, the reader is referred to Fishburn (1970), Kreps (1988), Gilboa (2009), Wakker (2010), and others.

Ellsberg Paradox

The classic Bayesian theory culminating in Savage's opus represents beliefs probabilistically, but it does not capture the degree of confidence that DMs have in their own probabilistic assessments, a degree that depends on the quality of the information that DMs use in forming these assessments. The classic theory focused on how to measure beliefs, without providing a way to assess the quality of such measurements.

Ellsberg (1961) provided two stark thought experiments that showed how this limitation may lead many people to violate Savage's otherwise extremely compelling axioms, and to express preferences that are incompatible with any (single, additive) probability measure. Ellsberg argued that a situation in which probabilities are not known, which he referred to as *ambiguity*,[8] induces different decisions than situations of risk, namely, uncertainty with known probabilities. Specifically, one of Ellsberg's experiments involves two urns, I and II, with 100 balls in each. The DM is told that

[6] Frisch (1926) was the first article we are aware of that adopted a similar approach in economic theory.

[7] See, e.g., chapter 8 of Kreps (1988).

[8] Today, the terms "ambiguity", "uncertainty" (as opposed to "risk"), and "Knightian uncertainty" are used interchangeably to describe the case of unknown probabilities.

(i) in both urns balls are either white or black;
(ii) in urn I there are 50 black and 50 white balls.

No information is given on the proportion of white and black balls in urn II. The DM has to choose both an urn and a color. A ball will be drawn at random from the urn that the DM named, and she will receive a prize if it is of the color stated.

The vast majority of decision makers are indifferent between betting on either color within each urn. However, not all are indifferent between the two urns. Specifically, many prefer either of the bets on the known urn (I) to either of the bets on the unknown urn (II). Clearly, no probability measure can justify such betting behavior.

Ellsberg's experiments revealed the phenomenon of *uncertainty aversion*, or *ambiguity aversion*: people tend to prefer situations with known probabilities to unknown ones, to the extent that these can be compared. Clearly, one can have the opposite phenomenon, of uncertainty/ambiguity liking, when people exhibit the opposite preferences. While gambling is an important exception, it is commonly assumed that people who are not uncertainty neutral tend to be uncertainty averse, in a way that parallels the common assumptions about attitudes toward risk.

Ellsberg's experiments are extremely elegant and they pinpoint precisely which of Savage's axioms is violated by DMs who are not indifferent between betting on the two urns.[9] But the elegance of these experiments is also misleading. Since they deal with balls and urns, and the information about the colors is completely symmetric, it is very tempting to adopt a probabilistic belief that would reflect this symmetry. Specifically, one may reason about the urn with unknown composition, "The number of red balls in it can be any number between 0 and 100. My information is completely symmetric, and there is no reason to believe that there are more red balls than black balls or vice versa. Hence, if I were to adopt a prior probability over the composition of the urn, from [0:100] to [100:0], I should choose a symmetric prior. That is, the probability that there are 3 red balls should be equal to the probability that there are 97 red balls, and so forth. In this case, the probability that a red ball is drawn out of the urn is precisely 50 %, and I should no longer express preferences for the known probabilities." Relatedly, one may also use the unknown urn to generate a bet with objective probability of 50 %: use an external chance device, which is known to be fair, and decide between betting on red or on black based on this device. If the DM has symmetric beliefs about the composition of the urn, she can thereby generate a bet that is equivalent to the bet on the urn with the known composition.

Based on such arguments, theorists often feel that there is no problem with subjective probabilities, at least as far as normative theories of choice are concerned. But this conclusion is wrong. In most real life examples there are no symmetries that allow the generation of risky bets. For example, suppose that Mary does not know what is the probability of the globe warming up by 4 degrees within the

[9] See Gilboa (2009) and Wakker (2010) for the analysis.

next 10 years. She cannot assume that this probability is 50 %, based on Laplace's Principle of Indifference (or "Principle of Insufficient Reason", Laplace 1814). The two eventualities, "average temperature increases by 4 degrees or more" and "average temperature does not increase by 4 degrees" are not symmetric. Moreover, if Mary replaces 4 degrees by 5 degrees, she will obtain two similar events, but she cannot generally assign a 50–50 % probability to any pair of complementary events. Nor will a uniform distribution over the temperature scale be a rational method of assigning probabilities.[10] The fundamental difficulty is that in most real life problems there is too much information to apply the Principle of Indifference, yet too little information to single out a unique probability measure.[11] Global warming and stock market crashes, wars and elections, business ventures and career paths face us with uncertainty that is neither readily quantified nor easily dismissed by symmetry considerations.

Other Disciplines

The Bayesian approach has proved useful in statistics, machine learning, philosophy of science, and other fields. In none of these fellow disciplines has it achieved the status of orthodoxy that it enjoys within economic theory. It is a respectable approach, providing fundamental insights and relishing conceptual coherence. It is worth pointing out, however, that in these disciplines the Bayesian approach is one among many. More importantly, in all of these disciplines the Bayesian approach is applied to a restricted state space, such as a space of parameters, whereas in economics it is often expected to apply also to a *grand state space*, whose elements describe anything that can possibly be of interest.

Consider statistics first. The statistical inference problem is defined by a set of distributions, or data generating processes, out of which a subset of distributions has to be chosen. In parametric problems, the set of distributions is assumed to be known up to the specification of finitely many parameters. Classical statistics does not allow the specification of prior beliefs over these parameters. By contrast, Bayesian

[10]Bertrand's (1907) early critique of the principle of indifference was made in the context of a continuous space. See also Gilboa (2009) and Gilboa et al. (2009).

[11]It is not entirely clear how one can justify the Principle of Indifference even in cases of ignorance. For example, Kass and Wasserman (1996) p. 1347 discuss the partition paradox and lack of parametric invariance, two closely related issues that arise with Laplace's Principle. Similar remarks from a Macroeconomics perspective can be found in Kocherlakota (2007) p. 357.

Based on a result by Henri Poincaré, Machina (2004) suggests a justification of the Laplace's Principle using a sequence of fine partitions of the state stace. This type of reasoning seems to underlie most convincing examples of random devices, such as tossing coins, spinning roulette wheels, and the like. It is tempting to suggest that this is the only compelling justification of the Principle of Indifference, and that this principle should not be invoked unless such a justification exists.

statistics demands that such beliefs be specified. Thus the Bayesian approach offers a richer language, within which the statistician can represent prior knowledge and intuition. Further, the Bayesian prior, updated to a posterior based on sampling, behaves in a much more coherent way than the techniques of classical statistics. (See, for example, Welch (1939), also described in DeGroot (1975), pp. 400–401.)

The main disadvantage of the Bayesian approach to statistics is its subjectivity: since the prior beliefs of the parameters is up to the statistician to choose, they will differ from one statistician to another. Admittedly, classical statistics cannot claim to be fully objective either, because the very formulation of the problem as well as the choice of statistics, tests, and significance levels leave room for the statistician's discretion. Yet, these are typically considered necessary evils, with objectivity remaining an accepted goal, whereas the Bayesian approach embraces subjective inputs unabashedly.[12] On the bright side, if a Bayesian statistician selects a sufficiently "diffused" or "uninformative" prior, she hopes not to rule out the true parameters a priori, and thereby to allow learning of objective truths in the long run, despite the initial reliance on subjective judgments.[13]

The Bayesian approach has a similar status in the related fields of computer science and machine learning.[14] On the one hand, it appears to be the most conceptually coherent model of inference. On the other, its conclusions depend on a priori biases. For example, the analysis of algorithms' complexity is typically conducted based on their worst case. The Bayesian alternative is often dismissed because of its dependence on the assumptions about the underlying distribution.

It is important to emphasize that in statistics and in computer science the state space, which is the subject of prior and posterior beliefs, tends to be a restricted space that does not grow with the data. For example, it can comprise of all combinations of values of finitely many parameters, which are held fixed throughout the sampling procedure. By contrast, the standard approach in economic theory suggests that the state of the world resolves all uncertainty, and thus describes everything that might be of relevance to the problem at hand, from the beginning of time until eternity. As a result, the state space that is often assumed in economics is much larger than in other disciplines. Importantly, it increases with the size of the data.

When one considers a restricted set of parameters, one may argue that the prior probability over this set is derived from past observations of similar problems, each with its own parameters, taken out of the same set. But when the grand state space is considered, and all past repetitions of the problem are already included in the

[12] See Lewis (1980) and chapter 4 of van Frassen (1989) (and the references therein) for a discussion of the relations between "objectivity" and subjective probabilities from a philosophical standpoint.

[13] Kass and Wasserman (1996), Bayarri and Berger (2004), and Berger (2004) discuss uninformative priors and related "objective" issues in Bayesian statistics (according to Efron (1986), some of these issues explain the relatively limited use of Bayesian methods in applied statistics).

[14] See Pearl (1986) and the ensuing literature on Bayesian networks.

description of each state, the prior probability should be specified on a rather large state space before any data were observed. With no observations at all, and a very large state space, the selection of a prior probability seems highly arbitrary.

In applications of the Bayesian approach in statistics, computer science, and machine learning, it is typically assumed that the basic structure of the process is known, and only a bounded number of parameters need to be learnt. Many non-parametric methods allow an infinitely dimensional parameter space, but one that does not grow with the number of observations. This approach is sufficient for many statistical inference and learning problems in which independent repetitions are allowed. But economics is often interested in events that do not repeat. Applying the Bayesian approach to these is harder to justify.

We are not fully aware of the origins of the application of the Bayesian approach to the grand state space. It is well known that de Finetti was a devout Bayesian. Savage, who followed his footsteps, was apparently much less religious in his Bayesian beliefs. Yet, he argued that a state of the world should "resolve all uncertainty" and, with a healthy degree of self-criticism, urged the reader to imagine that she had but one decision to be taken in her lifetime, and this is her choice of her strategy before being born. Harsanyi (1967, 1968) made a fundamental contribution to economics by showing how players' types should be viewed as part of the state of the world, and assumed that all unborn players start with a common prior over the grand state space that is thus generated. Aumann (1974, 1976, 1987) pushed this line further by assuming that all acts and all beliefs are fully specified in each and every state, while retaining the assumption that all players have a prior, and moreover, the same prior over the resulting state space.

Somewhere along recent history, with path-breaking contributions by de Finetti, Savage, Harsanyi, and Aumann, economic theory found itself with a state space that is much larger than anything that statisticians or computer scientists have in mind when they generate a prior probability. Surprisingly, the economic theory approach is even more idealized than the Bayesian approach in the philosophy of science. There is nothing wrong in formulating the grand state space as a canonical model within which claims can be embedded. But the assumption that one can have a prior probability over this space, or that this is the only rational way to think about it is questionable.

Summary

Since the mid-twentieth century economic theory has adopted a rather unique commitment to the Bayesian approach. By and large, the Bayesian approach is assumed to be the only rational way to describe knowledge and beliefs, and this holds irrespective of the state space under consideration. Importantly, economic theory clings to Bayesianism also when dealing with problems of unique nature, where nothing is known about the structure of the data generating process. Research in recent decades plainly shows that the Bayesian approach can be extremely

fruitful even when applied to such unique problems. But it is also possible that the commitment to the Bayesian approach beclouds interesting findings and new insights.

The preceding discussion highlights our view that there is nothing irrational about violating the Bayesian doctrine in certain problems. As opposed to models of bounded rationality, psychological biases, or behavioral economics, the focus of this survey are models in which DMs may sometimes admit that they do not know what the probabilities they face are. Being able to admit ignorance is not a mistake. It is, we claim, more rational than to pretend that one knows what cannot be known.

Bounded rationality and behavioral economics models often focus on descriptive interpretations. At times, they would take a conditionally-normative approach, asking normative questions given certain constraints on the rationality of some individuals. Such models are important and useful. However, the models discussed here are different in that they are fully compatible with normative interpretations. When central bank executives consider monetary policies, and when leaders of a country make decisions about military actions, they will not make a mistake if they do not form Bayesian probabilities. On the contrary, they will be well advised to take into account those uncertainties that cannot be quantified.

Alternative Models

The Anscombe-Aumann Setup

Anscombe and Aumann (1963) developed a version of the subjective expected utility model of Savage that turned out to be especially well suited for subsequent extensions of the basic Bayesian decision model. For this reason, in this sub-section we present this important setup.

The basic feature of the Anscombe-Aumann (AA, for short) model is that acts map states into lotteries, that is, acts' consequences involve exogenous probabilities a la von Neumann-Morgenstern. This feature is important both conceptually and mathematically. We now turn to introduce the setting formally, in the version presented by Fishburn (1970).

The set of simple probabilities $\Delta(X)$ on some underlying space X of alternatives is the space of consequences considered by the AA model.[15] There is a space of states of the world S endowed with an event algebra Σ. The objects of choice are acts, which map states into lotteries. We denote by $\mathcal{F}$ the collection of all simple acts $f : S \to \Delta(X)$, that is, acts that are finitely valued and Σ-measurable.[16]

[15] Throughout the section we use interchangeably the terms lotteries and simple probabilities.

[16] Simple acts have the form $f = \sum_{i=1}^{n} p_i 1_{E_i}$, where $\{E_i\}_{i=1}^{n} \subseteq \Sigma$ is a partition of S and $\{p_i\}_{i=1}^{n} \subseteq \Delta(X)$ is a collection of lotteries.

A key feature of $\Delta(X)$ is its convexity, which makes it possible to combine acts. Specifically, given any $\alpha \in [0,1]$, set

$$(\alpha f + (1-\alpha) g)(s) = \alpha f(s) + (1-\alpha) g(s), \qquad \forall s \in S. \tag{21.1}$$

The mixed act $\alpha f + (1-\alpha) g$ delivers in each state s the compound lottery $\alpha f(s) + (1-\alpha) g(s)$. In other words, *ex post*, after the realization of state s, the DM obtains a risky outcome governed by the lottery $\alpha f(s) + (1-\alpha) g(s)$.[17]

The possibility of mixing acts is a key dividend of the assumption that $\Delta(X)$ is the consequence space, which gives the AA setting a vector structure that the Savage setting did not have. The derivation of the subjective expected utility representation in the AA setting is based on this vector structure.

Risk preference The DM has a primitive preference $\succsim$ on $\mathcal{F}$. In turn, this preference induces a preference $\succsim_\Delta$ on lotteries by setting, for all $p, q \in \Delta(X)$,

$$p \succsim_\Delta q \Leftrightarrow f \succsim g,$$

where f and g are the constant acts such that $f(s) = p$ and $g(s) = q$ for all $s \in S$.

Constant acts are not affected by state uncertainty, only by the risk due to the lotteries' exogenous probabilities. For this reason, $\succsim_\Delta$ can be seen as the risk preference of the DM. This is an important conceptual implication of having $\Delta(X)$ as the consequence space. This richer consequence space mathematically delivers a most useful vector structure, while from a decision theoretic standpoint it enriches the setting with a risk preference that allows to consider the DMs' risk behavior separately. Differently put, the AA consequence space can be viewed as derived from an underlying consequence space X a la Savage, enriched by a lottery structure that allows to calibrate risk preferences.

Alternatively, one may view AA's model as an improved version of de Finetti's (1931, 1937) axiomatic derivation of expected value maximization with subjective probabilities. de Finetti assumed additivity or linearity in payoffs. This is a problematic assumption if payoffs are monetary, but it is more palatable if payoffs are probabilities of receiving a fixed desirable outcome. Replacing the payoffs in de Finetti's model by probabilities of outcomes, one obtains a model akin to AA's.

In a sense, the AA model is a hybrid between vNM's and Savage's. Mathematically it is akin to the former, as it starts with a vNM theorem on a particular mixture space, and imposes additional axioms to derive subjective probabilities. Conceptually, it is closer to Savage's model, as it derives probabilities from preferences. Many view this derivation as conceptually less satisfactory than Savage's, because the latter does not assume probabilities, or any numbers for that matter, to be part of the data. Anscombe and Aumann, however, viewed the use of objective probabilities as a merit, because they believed that people think in terms of subjective probabilities

[17] For this reason, mixing acts in this way is sometimes called "ex post randomization." For recent models with ex ante randomization, see Epstein et al. (2007), Ergin and Sarver (2009), Seo (2009), and Saito (2015).

after they have internalized the concept of objective probability. Be that as it may, there is no doubt that the AA model has become the main testbed for new models of decision under uncertainty.[18]

Axioms We now make a few assumptions on the primitive preference $\succsim$. The first one is a standard weak order axiom.

AA.1 WEAK ORDER: $\succsim$ on $\mathcal{F}$ is complete and transitive.

The next axiom is a monotonicity assumption: if state by state an act f delivers a weakly better (risky) consequence than an act g, then f should be weakly preferred to g. It is a basic rationality axiom.

AA.2 MONOTONICITY: for any $f, g \in \mathcal{F}$, if $f(s) \succsim_{\Delta} g(s)$ for each $s \in S$, then $f \succsim g$.

Next we have an independence axiom, which is peculiar to the AA setting since it relies on its vector structure.

AA.3 INDEPENDENCE: for any three acts $f, g, h \in \mathcal{F}$ and any $0 < \alpha < 1$, we have

$$f \succ g \Rightarrow \alpha f + (1-\alpha)h \succ \alpha g + (1-\alpha)h. \tag{21.2}$$

According to this axiom, the DM's preference over two acts f and g is not affected by mixing them with a common act h. In the special case when all these acts are constant, axiom AA.3 reduces to von Neumann-Morgenstern's original independence axiom on lotteries.

We close with standard Archimedean and nontriviality assumptions.[19]

AA.4 ARCHIMEDEAN: let f, g, and h be any three acts in $\mathcal{F}$ such that $f \succ g \succ h$. Then, there are $\alpha, \beta \in (0,1)$ such that $\alpha f + (1-\alpha)h \succ g \succ \beta f + (1-\beta)h$.

AA.5 NONDEGENERACY: there are $f, g \in \mathcal{F}$ such that $f \succ g$.

We can now state the AA subjective expected utility theorem.

Theorem 1. *Let $\succsim$ be a preference defined on $\mathcal{F}$. The following conditions are equivalent:*

(i) $\succsim$ satisfies axioms AA.1–AA.5;

(ii) there exists a non-constant function $u : X \to \mathbb{R}$ and a probability measure $P : \Sigma \to [0,1]$ such that, for all $f, g \in \mathcal{F}$, $f \succsim g$ if and only if

$$\int_S \left(\sum_{x \in \operatorname{supp} f(s)} u(x) f(s) \right) dP(s) \geq \int_S \left(\sum_{x \in \operatorname{supp} g(s)} u(x) g(s) \right) dP(s). \tag{21.3}$$

[18] See Ghirardato et al. (2003) for a subjective underpinning of the AA setup.

[19] See Gilboa (2009) for some more details on them.

Moreover, P is unique and u is cardinally unique.[20]

The preference functional $V : \mathcal{F} \to \mathbb{R}$ in (21.3) has the form

$$V(f) = \int_S \left(\sum_{x \in \text{supp} f(s)} u(x) f(s) \right) dP(s) \tag{21.4}$$

and consists of two parts. The inner part

$$\sum_{x \in \text{supp} f(s)} u(x) f(s) \tag{21.5}$$

is the expected utility of the lottery $f(s)$ that act f delivers when state s obtains. It is easy to see that this expected utility represents the DM's risk preference $\succsim_\Delta$. The outer part

$$\int_S \left(\sum_{x \in \text{supp} f(s)} u(x) f(s) \right) dP(s)$$

averages all expected utilities (21.5) according to the probability P, which quantifies the DM's beliefs over the state space.

The classical models of Savage and Anscombe-Aumann were considered the gold standard of decision under uncertainty, despite the challenge posed by Ellsberg's experiments. In the 1980s, however, several alternatives were proposed, most notably models based on probabilities that are not necessarily additive, or on sets of probabilities. We now turn to review these contributions and some of the current research in the area.

Choquet Expected Utility

The first general-purpose, axiomatically-based non-Bayesian decision model was the Choquet Expected Utility (CEU) model proposed by David Schmeidler in 1982, which appeared as Schmeidler (1989). Schmeidler's starting point was that the Bayesian model is a straightjacket that does not allow the DM to express her own degree of confidence in her beliefs. Schmeidler gave the example of two coins, one that has been tested extensively and is known to be fair, and the other about which nothing is known. He noted that a Bayesian would probably have 50–50 % beliefs regarding the result of the toss of either coin, but that these beliefs differ:

[20]Throughout the paper, cardinally unique means unique up to positive affine transformations.

in one case, the DM practically knows that each side of the coin has probability of 50 % of coming up. In the other case, the numbers 50–50 % are obtained with a shrug of one's shoulders, relying on symmetry of ignorance rather than symmetry of information.[21] Observe that Schmeidler's two-coin example is very close to Ellsberg's two-urn experiment. However, Schmeidler was not motivated by the desire to explain Ellsberg's results; rather, he considered the standard theory and found it counter-intuitive.

Schmeidler (1989) suggested to model probabilities by set functions that are not necessarily additive. For example, if H (T) designates the event "the unknown coin falls with H (T) up", and ν is the measure of credence, we may have

$$\nu(H) + \nu(T) < \nu(H \cup T)$$

Thus, the "probability" of events, as measured by our willingness to bet on them, may not satisfy the standard axioms of probability theory. Schmeidler referred to them as *non-additive probabilities*, and required that they be positive and monotone with respect to set inclusion. Such mathematical entities are also known by the term *capacities*. Formally, given an event algebra Σ of state space S, a set function $\nu : \Sigma \to [0, 1]$ is a capacity if

(i) $\nu(\emptyset) = 0$ and $\nu(S) = 1$;
(ii) $E \subseteq E'$ implies $\nu(E) \leq \nu(E')$.

Dempster (1967) and Shafer (1976) also suggested a theory of belief in which the degree of belief in an event did not obey additivity. They focused on the representation of uncertainty by *belief functions*. There is a vast literature that followed, often referred to as "imprecise probabilities" (see Walley 1991). Most of this literature, however, does not address the question of decision making. By contrast, Schmeidler had decision theory in mind, and he sought a notion of integration that would generalize standard expectation when the capacity ν happens to be additive. Such a notion of integration was suggested by Choquet (1953).

To understand the gist of the Choquet integral,[22] suppose that Σ is a σ-algebra (e.g., the power set 2^S) and consider a positive and bounded Σ-measurable function $\phi : S \to \mathbb{R}$. The *Choquet integral* of ϕ with respect to a capacity ν is given by:

$$\int \phi \, d\nu = \int_0^\infty \nu\left(\{s \in S : \phi(s) \geq t\}\right) dt, \tag{21.6}$$

where on the right-hand side we have a Riemann integral. To see why the Riemann integral is well defined, observe that the sets $E_t = \{s \in S : \phi(s) \geq t\}$ define a

[21] See Fischhoff and Bruine De Bruin (1999) for experimental evidence on how people use 50–50 % statements in this sense.

[22] We refer the interested reader to Denneberg (1994) and to Marinacci and Montrucchio (2004) for detailed expositions of Choquet integration.

chain that is decreasing in t (in the sense of set inclusion), and, since a capacity, is monotone, $\nu(E_t)$ is a decreasing function of t. For more detailed explanation of the Choquet integral the reader is referred to Gilboa (2009).

Schmeidler (1989) axiomatized Choquet expected utility in the AA setup. The key innovation relative to the AA axioms AA.1–AA.4 was to restrict the Independence axiom AA.3 to *comonotonic* acts, that is, acts $f, g \in \mathcal{F}$ for which it is never the case that both $f(s) \succ f(s')$ and $g(s) \prec g(s')$ for some states of the world s and s'. This is the preference version of comonotonicity.

S.3 COMONOTONIC INDEPENDENCE: for any pairwise comonotonic acts $f, g, h \in \mathcal{F}$ and any $0 < \alpha < 1$,

$$f \succ g \Rightarrow \alpha f + (1-\alpha)h \succ \alpha g + (1-\alpha)h. \tag{21.7}$$

According to this axiom, the DM's preference between two comonotonic acts f and g is not affected by mixing them with another act h that is comonotonic with both. The intuition behind this axiom can best be explained by observing that the classical independence axiom may not be very compelling in the presence of uncertainty. For example, assume that there are two states of the world, and two vNM lotteries $P \succ Q$. Let $f = (P, Q)$ and $g = (Q, P)$. Suppose that, due to ignorance about the state of the world, the DM is driven to express indifference, $f \sim g$. By AA's independence, for every h we will observe

$$\frac{1}{2}f + \frac{1}{2}h \sim \frac{1}{2}g + \frac{1}{2}h$$

However, for $h = g$ this implies that $\frac{1}{2}f + \frac{1}{2}g \sim g$, despite the fact that the act $\frac{1}{2}f + \frac{1}{2}g$ is risky while g is uncertain.

In this example, g can serve as a hedge against the uncertainty inherent in f, but it clearly cannot hedge against itself. The standard independence axiom is too demanding, because it does not distinguish between mixing operations $\alpha f + (1-\alpha)h$ that reduce uncertainty (via hedging) and mixing operations that do not. Restricting the independence axiom to pairwise comonotonic acts neutralizes this asymmetric effect of hedging.

Using the Comonotonic Independence axiom S.3, Schmeidler (1989) was able to prove the following representation theorem, which generalizes the subjective expected utility representation established by Theorem 1 by allowing for possibly non-additive probabilities. The proof of the result is based on some results on Choquet integration established in Schmeidler (1986).

Theorem 2 ((ii)). *Let $\succsim$ be a preference defined on $\mathcal{F}$. The following conditions are equivalent:*

(i) $\succsim$ satisfies axioms AA.1, AA.2, S.3 (Comonotonic Independence), AA.4, and AA.5;

(ii) there exists a non-constant function $u : X \to \mathbb{R}$ *and a capacity* $\nu : \Sigma \to [0,1]$ *such that, for all* $f, g \in \mathcal{F}$, $f \succsim g$ *if and only if*

$$\int_S \left(\sum_{x \in \text{supp} f(s)} u(x) f(s) \right) d\nu(s) \geq \int_S \left(\sum_{x \in \text{supp}\, g(s)} u(x) g(s) \right) d\nu(s). \tag{21.8}$$

Moreover, ν *is unique and* u *is cardinally unique.*

Gilboa (1987), Wakker (1989a,b), and Nakamura (1990) established purely subjective versions of Schmeidler's representation result.[23] Sarin and Wakker (1992) showed that the existence of a suitable rich collection of unambiguous events substantially streamlines the derivation of Schmeidler's representation through a simple cumulative dominance condition.

Maxmin Expected Utility

Schmeidler's model is a generalization of Anscombe-Aumann's in a way that allows us to cope with uncertainty, or ambiguity. The capacity in the model can be interpreted as a lower bound on probabilities. Specifically, let $\Delta(\Sigma)$ be the collection of all finitely additive probability measures $P : \Sigma \to [0,1]$ and define the *core* of ν to be, as in cooperative game theory,

$$core(\nu) = \{P \in \Delta(\Sigma) : P(E) \geq \nu(E) \text{ for all } E \in \Sigma\}.$$

If $core(\nu) \neq \emptyset$, we may think of $\nu(E)$ as the lower bound on $P(E)$, and then ν is a concise way to represent a set of probabilities, presumably those that are considered possible. The lower envelope of a set of probabilities is also the common interpretation of belief functions (Dempster 1967; Shafer 1976).

Schmeidler (1986) has shown that if ν is *convex* in the sense that

$$\nu(E) + \nu(E') \leq \nu(E \cup E') + \nu(E \cap E'), \qquad \forall E, E' \in \Sigma,$$

(Shapley 1972) then

$$\int_S \phi \, d\nu = \min_{P \in core(\nu)} \int_S \phi \, dP \tag{21.9}$$

for every Σ-measurable bounded function $\phi : S \to \mathbb{R}$ (see also Rosenmueller 1971, 1972). Thus, when the capacity ν happens to be convex (e.g., a belief function a la

[23] Nakamura and Wakker's papers use versions of the so-called *tradeoff method* (see Kobberling and Wakker (2003), for a detailed study of this method and its use in the establishment of axiomatic foundations for choice models).

Dempster-Shafer), Choquet integration has a simple and intuitive interpretation: a DM who evaluated an act f by the Choquet integral of its utility profile $u \circ f$ can be viewed as if she entertained a *set* of possible probabilities, $core(\nu)$, and evaluated each act by its minimal expected utility, over all probabilities in the set.

There is a simple behavioral condition that characterizes CEU preferences with convex ν. To introduce it, denote by $B_0(\Sigma)$ the vector space of all simple functions $\phi : S \to \mathbb{R}$ and consider the Choquet functional $I : B_0(\Sigma) \to \mathbb{R}$ given by $I(\phi) = \int \phi \, d\nu$. This functional is easily seen to be concave when (21.9) holds. Actually, according to a classic result of Choquet (1953), I is concave if and only if its capacity ν is convex.[24] This concavity property suggests the following convexity axiom, due to Schmeidler (1989), which models a negative attitude toward ambiguity.

S.6 UNCERTAINTY AVERSION: for any $f, g \in \mathcal{F}$ and any $0 < \alpha < 1$, we have

$$f \sim g \Rightarrow \alpha f + (1-\alpha) g \succsim f.$$

Thus, uncertainty aversion states that mixing, through randomization, between equivalent acts can only make the DM better off. For example, in Ellsberg's example it is natural to expect that DMs prefer to hedge against ambiguity by mixing acts *IIB* and *IIW*, that is,

$$\alpha IIB + (1-\alpha) IIW \succsim IIB \sim IIW, \qquad \forall \alpha \in [0,1].$$

This mixing can be thus viewed as a form of hedging against ambiguity that the DM can choose.[25]

Theorem 3. *In Theorem 2, $\succsim$ satisfies axiom S.6 if and only if the capacity ν in (21.8) is convex.*

Theorem 3 (Schmeidler 1989) shows that convex capacities characterize ambiguity averse Choquet expected utility DMs (in the sense of axiom S.6). Since most DMs are arguably ambiguity averse, this is an important result in Choquet expected utility theory. Moreover, relating this theory to maximization of the worst-case expected utility over a set of probabilities has several advantages. First, it obviates the need to understand the unfamiliar concept of Choquet integration. Second, it provides a rather intuitive, if extreme, cognitive account of the decision process: as in classical statistics, the DM entertains several probability measures as potential beliefs. Each such "belief" induces an expected utility index for each act. Thus, each act has many expected utility values. In the absence of second-order beliefs, the cautious DM chooses the worst-case expected utility as summarizing the act's desirability. Wakker (1990, 1991) established several important behavioral properties and characterizations of concave/convex capacities in the CEU model.

[24] See Marinacci and Montrucchio (2004) p. 73. They show on p. 78 that (21.9) can be derived from this result of Choquet through a suitable application of the Hahn-Banach Theorem.

[25] Klibanoff (2001a,b) studied in detail the relations between randomization and ambiguity aversion.

Gilboa and Schmeidler (1989) This account of Choquet expected utility maximization also relates to the maxmin criterion of Wald (1950; see also Milnor 1954). However, there are many natural sets of probabilities that are not the core of any capacity. Assume, for example, that there are three states of the world, $S = \{1, 2, 3\}$. Assume that the DM is told that, if state 1 is not the case, then the (conditional) probability of state 2 is at least $2/3$. If this is all the information available to her, she knows only that state 2 is at least twice as likely than state 3. Hence the set of probability vectors $P = (p_1, p_2, p_3)$ that reflects the DM's knowledge consists of all vectors such that

$$p_2 \geq 2p_3$$

It is easy to verify that this set is not the core of a capacity. Similarly, one may consider a DM who has a certain probability measure P in mind, but allows for the possibility of error in its specification. Such a DM may consider a set of probabilities

$$C = \{Q \in \Delta(\Sigma) : \|P - Q\| < \varepsilon\}$$

for some norm $\|\cdot\|$ and $\varepsilon > 0$, and this set is not the core of any capacity (such sets were used in Nishimura and Ozaki (2007)).

It therefore makes sense to generalize Choquet expected utility with convex capacities to the maxmin rule, where the minimum is taken over general sets of probabilities. Decision rules of this type have been suggested first by Hurwicz (1951), under the name of Generalized Bayes-minimax principle, and then by Smith (1961), Levi (1974, 1980), and Gärdenfors and Sahlin (1982). Recently, related ideas appeared in mathematical finance (see Artzner et al. 1997, 1999).

Gilboa and Schmeidler (1989) provided an axiomatic model of maxmin expected utility maximization ("MMEU", also referred to as "MEU"). This model is also formulated in the AA framework and, like the Choquet expected utility model, is based on a suitable weakening of the Independence axiom AA.3. Schmeidler's Comonotonic Independence axiom restricted AA.3 to the case that all acts are pairwise comonotonic. This rules out obvious cases of hedging, but it may allow for more subtle ways in which expected utility can be "smoothed out" across states of the world.[26] A more modest requirement restricts the independence condition to the case in which the act h is constant:

[26]For example, assume that there are three states of the world, and two acts offer the following expected utility profiles: $f = (0, 10, 20)$ and $g = (4, 10, 14)$. Assume that the DM is indifferent between f and g, that is, that she is willing to give up 1 unit of expected utility in state 3 in order to transfer 5 units from state 3 to state 1. Comonotonic independence would imply that the DM should also be indifferent between f and g when they are mixed with any other act comonotonic with both, such as f itself. However, while f clearly doesn't offer a hedge against itself, mixing f with g can be viewed as reducing the volatility of the latter, resulting in a mix that is strictly better than f and g.

GS.3 C-INDEPENDENCE: for all acts $f, g \in \mathcal{F}$ and all lottery acts p,

$$f \succsim g \Rightarrow \alpha f + (1-\alpha)p \succsim \alpha g + (1-\alpha)p, \qquad \forall \alpha \in [0,1].$$

C-Independence is essentially weaker than Comonotonic Independence S.3 because lottery (constant) acts are comonotonic with all other acts.[27] The axiom is arguably easier to accept because the mixture with a lottery act can be viewed as a change of the unit of measurement. Indeed, this axiom may be viewed as the preference version of the following property of real-valued functionals: a functional $I : B_0(\Sigma) \to \mathbb{R}$ is said to be *translation invariant* if

$$I(\alpha\phi + k) = \alpha I(\phi) + I(k), \qquad \forall \alpha \in \mathbb{R},$$

given any $\phi \in B_0(\Sigma)$ and any constant function k.[28]

Gilboa and Schmeidler thus used a weaker version of the independence axiom, but they also imposed the uncertainty aversion axiom S.6. Both axioms GS.3 (C-Independence) and S.6 (Uncertainty Aversion) follow from the Independence axiom AA.3. Thus, the following representation result, due to Gilboa and Schmeidler (1989), generalizes Theorem 1 by allowing for possibly nonsingleton sets of probabilities.

Theorem 4. *Let $\succsim$ be a preference defined on $\mathcal{F}$. The following conditions are equivalent:*

(i) $\succsim$ satisfies axioms AA.1, AA.2, GS.3 (C-Independence), AA.4, AA.5, and S.6 (Uncertainty Aversion);

(ii) there exists a non-constant function $u : X \to \mathbb{R}$ and a convex and compact set $C \subseteq \Delta(\Sigma)$ of probability measures such that, for all $f, g \in \mathcal{F}$,

$$\begin{aligned} f \succsim g \Leftrightarrow \min_{P \in C} \int_S \left(\sum_{x \in \text{supp} f(s)} u(x) f(s) \right) dP(s) \\ \geq \min_{P \in C} \int_S \left(\sum_{x \in \text{supp} f(s)} u(x) g(s) \right) dP(s), \end{aligned} \tag{21.10}$$

Moreover, C is unique and u is cardinally unique.

[27] Schmeidler required that all three acts be pairwise comonotonic, whereas C-Independence does not restrict attention to comonotonic pairs (f, g). Thus, C-Independence is not, strictly speaking, weaker than Comonotonic Independence. However, in the presence of Schmeidler's other axioms, Comonotonic Independence is equivalent to the version in which f and g are not required to be comonotonic.

[28] See Ghirardato et al. (1998) for details.

The set C is a singleton if and only if $\succsim$ satisfies the Independence axiom AA.3. A slightly more interesting result actually holds, which shows that maxmin expected utility DMs reduce to subjective expected utility ones when their choices do not involve any hedging against ambiguity.[29]

Proposition 5. *In Theorem 4, C is a singleton if and only if, for all $f, g \in \mathcal{F}$,*

$$f \sim g \Rightarrow \frac{1}{2}f + \frac{1}{2}g \sim g.$$

When C is not a singleton, the model can express more complex states of knowledge, reflected by various sets C of probabilities. For applications in economic theory, the richness of the maxmin model seems to be important. In particular, one may consider any model in economic theory and enrich it by adding some uncertainty about several of its parameters. By contrast, in order to formulate Choquet expected utility, one needs to explicitly consider the state space and the capacity defined on it. Often, this exercise may be intractable.

By contrast, for some practical applications such as in medical decision making, the richness of the maxmin model may prove a hindrance. Wakker (2010) presents the theory of decision making under risk and under ambiguity geared for such applications. He focuses on capacities as a way to capture ambiguity, rather than on sets of probabilities.[30]

The maxmin model allows for more degrees of freedom than the CEU model, but it does not generalize it. In fact, the overlap of the two models is described in Theorem 3 and occurs when the uncertainty averse axiom S.6 holds. But, whereas uncertainty aversion – through axiom S.6 – is built into the decision rule of the maxmin model, Choquet expected utility can express attitudes of uncertainty liking. This observation in part motivated the search by Ghirardato et al. (2004) of a class of preferences that may not satisfy S.6 and is able to encompass both CEU and MMEU preferences. We review this contribution below.

Finally, Casadesus-Masanell et al. (2000), Ghirardato et al. (2003), and Alon and Schmeidler (2014) established purely subjective versions of Gilboa and Schmeidler's representation result.[31]

Countably additive priors Theorem 4 considers the set $\Delta(\Sigma)$ of all finitely additive probabilities. In applications, however, it is often important to consider

[29] See Ghirardato et al. (2004) for details.

[30] Wakker (2010) also introduces the gain-loss asymmetry that is one of the hallmarks of Prospect Theory (Kahneman and Tversky 1979). The combination of gain-loss asymmtry with rank-dependent expected utility (Quiggin 1982; Yaari 1987) resulted in Cumulative Prospect Theory (CPT, Tversky and Kahneman 1992). When CPT is interpreted as dealing with ambiguity, it is equivalent to Choquet expected utility with the additional refinement of distinguishing gains from losses.

[31] For a critical review of the maxmin and other non-Bayesian models, see Al-Najjar and Weinstein (2009) (see Mukerji 2009; Siniscalchi 2009b, for a discussion).

countably additive probabilities, which have very convenient analytical properties that many important results in probability theory crucially rely upon.

The behavioral condition that underlies countably additive priors is Monotone Continuity, introduced by Arrow (1970) to characterize countable additivity of the subjective probability P in Savage's model.

MC MONOTONE CONTINUITY: If $f, g \in \mathcal{F}$, $x \in X$, $\{E_n\}_{n\geq 1} \in \Sigma$ with $E_1 \supseteq E_2 \supseteq \ldots$ and $\bigcap_{n\geq 1} E_n = \emptyset$, then $f \succ g$ implies that there exists $n_0 \geq 1$ such that $xE_{n_0}f \succ g$.

Marinacci (2002a) and Chateauneuf et al. (2005) showed that this condition keeps characterizing countable additivity in the MMEU model. Next we state a version of their results, a countably additive counterpart of Theorem 4. Here $Q \ll P$ means that Q is absolutely continuous with respect to P, i.e., $P(E) = 0$ implies $Q(E) = 0$ for all $E \in \Sigma$.

Theorem 6. *In Theorem 4, $\succsim$ satisfies Axiom MC if and only if all probabilities in C are countably additive. In this case, there exists $P \in C$ such that $Q \ll P$ for all $Q \in C$.*

Besides the countable additivity of priors, axiom MC also delivers the existence of a "control" prior $P \in C$ relative to which all other priors $Q \in C$ are absolutely continuous.[32]

In decision theory the use of countably additive priors has been often debated, most forcefully by de Finetti and Savage themselves, who argued that it is a purely technical property that, if anything, actually impairs the analysis (e.g., over countable state spaces it is not possible to define uniform priors that are countably additive). However, Arrow's characterization of countably additive priors in Savage's model through Monotone Continuity and its MMEU version in Theorem 6 show that behaviorally this technically most useful property requires a relatively small extra baggage compared to the basic axioms of the finitely additive case.[33]

Equivalent priors A minimal consistency requirement among priors in C is that they agree on what is possible or impossible. Formally, this is the case if any two priors P and P' in C are *equivalent*, i.e., if they are mutually absolutely continuous

[32] As Chateauneuf et al. (2005) show, this control prior exists because, under Axiom MC, the set C is weakly compact, a stronger compactness condition than the weak*-compactness that C features in Theorem 4. Their results have been generalized to variational preferences by Maccheroni et al. (2006a).

[33] In this regard, Arrow (1970) wrote that "the assumption of Monotone Continuity seems, I believe correctly, to be the harmless simplification almost inevitable in the formalization of any real-life problem." See Kopylov (2010) for a recent version of Savage's model under Monotone Continuity.

In many applications, countable additivity of the measure(s) necessitates the restriction of the algebra of events to be a proper subset of 2^S. Ignoring many events as "non-measurable" may appear as sweeping the continuity problem under the measurability rug. However, this approach may be more natural if one does not start with the state space S as primitive, but derives it as the semantic model of a syntactic system, where propositions are primitive.

($P(E) = 0$ if and only if $P'(E) = 0$ for all $E \in \Sigma$). Epstein and Marinacci (2007) provide a behavioral condition that ensures this minimal consistency among priors, which is especially important in dynamic problems that involve priors' updating.

Interestingly, this condition turns out to be a translation in a choice under uncertainty setup of a classic axiom introduced by Kreps (1979) in his seminal work on menu choices. Given any two consequences x and y, let

$$x \vee y = \begin{cases} x \text{ if } x \succsim y \\ y \text{ otherwise} \end{cases}$$

and given any two acts f and g, define the act $f \vee g$ by $(f \vee g)(s) = f(s) \vee g(s)$ for each $s \in S$.

GK GENERALIZED KREPS: For all $f, f', g \in \mathcal{F}$, $f \sim f \vee f' \Rightarrow f \vee g \sim (f \vee g) \vee f'$.

In every state, the act $f \vee f'$ gives the better of the two outcomes associated with f and f'. Thus we say that $f \vee f'$ weakly improves f in 'the direction' f'. GK requires that if an improvement of f in direction f' has no value, then the same must be true for an improvement in direction f' of any act (here $f \vee g$) that improves f. The next result of Epstein and Marinacci (2007) shows that for maxmin preferences this seemingly innocuous axiom is equivalent to the mutual absolute continuity of priors.

Theorem 7. *In Theorem 4, $\succsim$ satisfies Axiom GK if and only if the probabilities in C are equivalent.*

Unanimity Preferences

Another way to deal with ambiguity is to relax the completeness of preferences. Indeed, because of the poor information that underlies ambiguity, the DM may not be able to rank some pairs of acts. If so, one of the most basic assumptions in decision theory, namely, that preferences are complete, may be relaxed because of ambiguity.

This is the approach proposed by Truman Bewley. Incomplete preferences were already studied by Aumann (1962), interpreted as a DM's inability to decide between some pairs of alternatives. Building on Aumann's work, Bewley presented in 1986 a model of incomplete preferences in the context of uncertainty, which appeared as Bewley (2002). In his model the Weak Order Axiom AA.1 is replaced by two weaker assumptions.

B.1a PARTIAL ORDER**:** $\succsim$ on $\mathcal{F}$ is reflexive and transitive.

Hence, $\succsim$ is no longer required to be complete. The DM, however, knows her tastes: the only reason for incompleteness is ignorance about probabilities. For this reason, Bewley assumes the next weak form of completeness, which only applies to lottery acts.

B.1b C-COMPLETENESS**:** for every lottery acts $p, q \in \Delta(X)$, $p \succsim q$ or $q \succsim p$.

In other words, B.1 requires the risk preference $\succsim_\Delta$ to be complete. Using these two axioms, Gilboa et al. (2010) established the following general form of Bewley's representation theorem.[34]

Theorem 8. *Let $\succsim$ be a preference defined on $\mathcal{F}$. The following conditions are equivalent:*

(i) $\succsim$ satisfies axioms B.1, and AA.2–AA.5;
(ii) there exists a non-constant function $u : X \to \mathbb{R}$ and a convex and compact set $C \subseteq \Delta(\Sigma)$ of probability measures such that, for all $f, g \in \mathcal{F}$,

$$f \succsim g \Leftrightarrow \int_S \left(\sum_{x \in \text{supp} f(s)} u(x) f(s) \right) dP(s)$$
$$\geq \int_S \left(\sum_{x \in \text{supp} f(s)} u(x) g(s) \right) dP(s), \qquad \forall P \in C. \tag{21.11}$$

Moreover, C is unique and u is cardinally unique.

In this representation a set of probability measures C arises, interpreted as the probabilistic models that are compatible with the DM's information. Two acts f and g are comparable only when their expected utilities with respect to the probabilities in C unanimously rank one act over the other. If this is not the case – that is, if the probabilities in C do not agree in ranking of the two acts– the DM is unable to rank the two acts.

When preferences are incomplete, the model does not always specify what the DM will do. In particular, acts are not evaluated by a numerical index V that represents preferences and that makes it possible to formulate the optimization problems that most economic applications feature. To complete the model, one needs to add some assumptions about choices in case preferences do not have a maximum. One possibility is to assume that there exists a status quo, namely, an alternative that remains the default choice unless it is dethroned by another alternative that is unanimously better. This might be a rather reasonable descriptive model, especially of organizations, but it is considered by many to be less than rational. Recently, Ortoleva (2010) reconsidered Bewley's inertia insight from a different angle by showing, within a full-fledged axiomatic model, how status quo biases may lead to incomplete preferences.

Another approach suggests to complete preferences based on the same set of probabilities C. Gilboa et al. (2010) offer a model involving two preference

[34] A caveat: the unanimity rule (21.11) is slightly different from Bewley's, who represents strict preference by unanimity of strict inequalities. This is generally not equivalent to representation of weak preference by unanimity of weak inequalities.

relations, and show that certain axioms, stated on each relation separately as well as relating the two, are equivalent to a joint representation of the two relations by the same set of probabilities C: one by the unanimity rule, and the other – by the maxmin rule. Their results provide a bridge between the two classic representations (21.10) and (21.11), as well as a possible account by which maxmin behavior might emerge from incomplete preferences.

Unanimity, Scenarios, and Uncertainty Aversion

Ghirardato et al. (GMM, 2004) used some insights from Bewley's unanimity representation to remove the Uncertainty Aversion axiom S.6 in the derivation of Gilboa and Schmeidler (1989) and, in this way, to propose a class of preferences that encompasses both Choquet and maxmin preferences. To this end, they consider the following definition.

Definition 9. A preference $\succsim$ on $\mathcal{F}$ is said to be invariant biseparable if it satisfies axioms AA.1, AA.2, GS.3 (C-Independence), AA.4, and AA.5.

Invariant biseparable (IB) preferences thus satisfy all AA axioms, except for the Independence axiom AA.3, which is replaced by the C-Independence axiom GS.3 of Gilboa and Schmeidler (1989).[35] Thanks to this key weakening, invariant biseparable preferences include as special cases both CEU and MMEU preferences: the former constitute the special case when the Comonotonic Independence Axiom S.3 holds, while the latter – when the Uncertainty Aversion axiom S.6 holds.

The main tool that GMM use to study IB preferences is an auxiliary relation $\succsim^*$ on $\mathcal{F}$. Specifically, given any two acts $f, g \in \mathcal{F}$, act f is said to be *unambiguously (weakly) preferred* to g, written $f \succsim^* g$, if

$$\alpha f + (1-\alpha)\, h \succsim \alpha g + (1-\alpha)\, h$$

for all $\alpha \in [0,1]$ and all $h \in \mathcal{F}$. In words, $f \succsim^* g$ holds when the DM does not find any possibility of hedging against or speculating on the ambiguity that she may perceive in comparing f and g. GMM argue that this DM's choice pattern reveals that ambiguity does not affect her preference between f and g, and this motivates the "unambiguously preferred" terminology.

The unambiguous preference relation is, in general, incomplete. This incompleteness is due to ambiguity

Lemma 10. *The following statements hold:*

(i) If $f \succsim^* g$, then $f \succsim g$.
(ii) $\succsim^*$ satisfies axioms B.1, AA.2, and AA.3

[35] The name biseparable originates in Ghirardato and Marinacci (2001, 2002), which we will discuss later.

(iii) $\succsim^*$ *is the maximal restriction of* $\succsim$ *satisfying the Independence axiom AA.3.*[36]

By (i) and (ii), the unambiguous preference $\succsim^*$ is a restriction of the primitive preference relation $\succsim$ that satisfies reflexivity, transitivity, monotonicity, and independence. By (iii), it is the maximal such restriction that satisfies independence.[37]

The next result proves, along the lines of the Bewley-type representation (21.11), that the unambiguous preference can be represented by a set of priors.

Proposition 11. *Let* $\succsim$ *be an IB preference on* $\mathcal{F}$*. Then, there exists a function* $u : X \to \mathbb{R}$ *and a convex and compact set* $C \subseteq \Delta(\Sigma)$ *of probability measures such that, for all* $f, g \in \mathcal{F}$*,*

$$f \succsim^* g \Leftrightarrow \int_S u(f)\, dP(s) \geq \int_S u(g)(s)\, dP(s), \qquad \forall P \in C. \tag{21.12}$$

In words, f is unambiguously weakly preferred to g if and only if every probability $P \in C$ assigns a weakly higher expected utility to f. It is natural to interpret each prior $P \in C$ as a "possible scenario" that the DM envisions, so that unambiguous preference corresponds to preference in every scenario. GMM thus argue that C represents the (subjective) perception of ambiguity of the DM, and that the DM perceives ambiguity in a decision problem if C is not a singleton.

The relation $\succsim^*$ thus makes it possible to elicit a set of priors C for a general IB preference $\succsim$. When $\succsim$ is a MMEU preference, C is the set of priors of the maxmin representation (21.10). When $\succsim$ is a CEU preference that satisfies axiom S.6, C is the core of the representing capacity ν.[38]

More generally, GMM prove a representation theorem for IB preferences based on the set C, which generalizes Theorems 2 and 4. To this end, given any act f consider its expected utility profile $\left\{\int_S u(f)\, dP(s) : P \in C\right\}$ under C. Write $f \asymp g$ if two acts f and g feature isotonic profiles, that is,

$$\int_S u(f(s))\, dP'(s) \geq \int_S u(f(s))\, dP''(s) \Leftrightarrow \int_S u(g(s))\, dP'(s)$$
$$\geq \int_S u(g(s))\, dP''(s), \qquad \forall P', P'' \in C.$$

Intuitively, in this case the DM perceives a similar ambiguity in both acts. For example, $p \asymp q$ for all lottery acts, which are unambiguous.

It is easy to see that $\asymp$ is an equivalence relation. Denote by $[f]$ the relative equivalence class determined by an act f, and by $\mathcal{F}_{|\asymp}$ the quotient space of $\mathcal{F}$ that consists of these equivalence classes.

[36] That is, if $\succsim' \subseteq \succsim$ and $\succsim'$ satisfies independence, then $\succsim' \subseteq \succsim^*$.

[37] This latter feature of $\succsim^*$ relates this notion to an earlier one by Nehring (2001), as GMM discuss.

[38] GMM also show the form that C takes for some CEU preferences that do not satisfy S.6.

Theorem 12. *Let $\succsim$ be an IB preference on $\mathcal{F}$. Then, there exists a function $u : X \to \mathbb{R}$, a convex and compact set $C \subseteq \Delta(\Sigma)$ of probability measures, and a function $a : \mathcal{F}_{|\succ} \to [0,1]$ such that $\succsim$ is represented by the preference functional $V : \mathcal{F} \to \mathbb{R}$ given by*

$$V(f) = a([f]) \min_{P \in C} \int_S u(f(s))\, dP(s) + (1 - a([f])) \max_{P \in C} \int_S u(f(s))\, dP(s), \tag{21.13}$$

where u and C represent $\succsim^*$ in the sense of (21.12).

Moreover, C is unique, u is cardinally unique, and a is unique on $\mathcal{F}_{|\succ}$ (with the exclusion of the equivalence class $[p]$ of lottery acts).

In this representation, the revealed perception of ambiguity, embodied by the set C, is separated from the DM's reaction to it, modelled by the function a. Both C and a are derived endogenously within the model. When a is constant equal to 1, we get back to the maxmin representation. Otherwise, we have a more general choice criterion that may well exhibit ambiguity loving (the polar case is, clearly, when a is a constant equal to 0).

Giraud (2005) and Amarante (2009) studied invariant biseparable preferences, with novel important insights. Amarante established an alternative characterization of IB preferences through the two stage form

$$V(f) = \int_\Delta \left(\int_S u(f(s))\, dP(s) \right) d\nu(P)$$

where ν is a capacity over the set of measures $\Delta = \Delta(\Sigma)$ on S. In a statistical decision theory vein, the capacity ν quantifies DM's beliefs over the possible models P. Giraud thoroughly studies a similar representation, motivated by the desire to incorporate probabilistic information in a choice under ambiguity framework.

Finally, Siniscalchi (2006a) investigates an interesting class of invariant biseparable preferences that satisfy a local no-hedging condition that gives preferences a piecewise structure that makes them SEU on each component (see Castagnoli et al. 2003, for a related representation).

***a*-MEU Preferences** In the special case when the function a is constant the representation (21.13) reduces to

$$V(f) = a \min_{P \in C} \int_S u(f(s))\, dP(s) + (1 - a) \max_{P \in C} \int_S u(f(s))\, dP(s). \tag{21.14}$$

This is the a-MEU criterion that Jaffray (1989) suggested to combine Hurwicz (1951)'s criterion (see also Arrow and Hurwicz 1972) with a maxmin approach. Intuitively, $a \in [0,1]$ measures the degree of the individual's pessimism, where $a = 1$ yields the maxmin expected utility model, and $a = 0$ – its dual, the maxmax expected utility model. However, this apparently natural idea turned out to be surprisingly tricky to formally pin down. GMM provided a specific axiom

that reduces the IB representation to (21.14), where C represent $\succsim^*$ in the sense of (21.12). Because of this latter clause, when $a \in (0, 1)$ it is not possible to take any pair u and C as a given and assume that the DMs' preferences are represented by the corresponding a-MEU criterion (21.14). In a nutshell, the issue is the uniqueness properties of C in (21.14), which are problematic when $a \in (0, 1)$. We refer the reader to GMM and to Eichberger et al. (2008, 2011) for more on this issue. (The latter paper shows that for finite state spaces the a-MEU axiomatized as a very special case of (21.13) by GMM only allows for $\alpha = 0$ or $\alpha = 1$).

Smooth Preferences

The MMEU model discussed above is often viewed as rather extreme: if, indeed, a set of probability measures C is stipulated, and each act f is mapped to a range of expected utility values, $\left\{\int_S u(f)dp \mid p \in C\right\}$, why should such an f be evaluated by the minimal value in this interval? This worst-case scenario approach seems almost paranoid: why should the DM assume that nature[39] will choose a probability as if to spite the DM? Isn't it more plausible to allow for other ways that summarize the interval by a single number?

The extreme nature of the maxmin model is not evident from the axiomatic derivation of the model. Indeed, this model is derived from Anscombe-Aumann's by relaxing their independence axiom in two ways: first, by restricting it to mixing with a constant act (h above) and, second, by assuming uncertainty aversion. These weaker axioms do not seem to reflect the apparently-paranoid attitude of the maxmin principle. A question then arises, how do these axioms give rise to such extreme uncertainty attitude?

In this context it is important to recall that the axiomatic derivation mentioned above is in the revealed preferences tradition, characterizing behavior that could be represented in a certain mathematical formula. An individual who satisfies the axioms can be thought of *as if* she entertained a set C of priors and maximized the minimal expected utility with respect to this set. Yet, this set of priors need not necessarily reflect the individual's knowledge. Rather, information and personal taste jointly determine the set C. Smaller sets may reflect both better information and a less averse uncertainty attitude. For example, an individual who bets on a flip of a coin and follows the expected utility axioms with respect to a probability $p = 0.5$ of "Head" may actually know that the probability p is 0.5, or she may have no clue about p but chooses the model $p = 0.5$ because she is insensitive to her ignorance about the true data generating process. Thus, information and attitude to uncertainty are inextricably intertwined in the set C. More generally, it is possible that the individual has objective information that the probability is in a set D, but

[39]Relations between ambiguity and games against nature are discussed in Hart et al. (1994), Maccheroni et al. (2006a,b), and Ozdenoren and Peck (2008).

behaves according to the maxmin expected utility rule with respect to a set $C \subset D$, reflecting her uncertainty attitude. This intuition has motivated the model of Gajdos et al. (2008) that axiomatically established the inclusion $C \subset D$ (some related ideas can be found in Wang (2003a) and Giraud (2005)).

If, however, the set of priors C is interpreted cognitively a la Wald, that is, as the set of probabilities that are consistent with objectively available information, one may consider alternatives to the maxmin rule that, under this Waldean interpretation, has an extreme nature. One approach to address this issue is to assume that the DM has a prior probability over the possible probability distributions in C. Thus, if $\Delta(\Sigma)$ is the space of all "first order" probability distributions (viewed as data generating processes), and μ is a "second order" prior probability over them, one can use μ to have an averaging of sorts over all expected utility values of an act f.

Clearly, the expectation of expectations is an expectation. Thus, if one uses μ to compute the expectation of the expected utility, there will exist a probability $\hat{p}$ on S, given by

$$\hat{p} = \int_{\Delta(\Sigma)} p d\mu$$

such that for every act f (and every utility function u)

$$\int_{\Delta(\Sigma)} \left(\int_S u(f) dp \right) d\mu = \int_S u(f) d\hat{p}$$

In this case, the new model cannot explain any new phenomena, as it reduces to the standard Bayesian model. However, if the DM uses a non-linear function to evaluate expected utility values, one may explain non-neutral attitudes to uncertainty. Specifically, assume that

$$\varphi : \mathbb{R} \to \mathbb{R}$$

is an increasing function, and an act f is evaluated by

$$V(f) = \int_{\Delta(\Sigma)} \varphi \left(\int_S u(f) dp \right) d\mu.$$

In this representation, μ is read as representing information (about the probability model p), whereas φ reflects attitude towards ambiguity, with a concave φ corresponding to ambiguity aversion, similarly to the way that concave utility represents risk aversion in the classical model of expected utility under risk. In this way we have a separation between ambiguity perception, an information feature modelled by μ and its support, and ambiguity attitude, a taste trait modelled by φ and its shape.

This decision rule has been axiomatized by Klibanoff et al. (2005). It has become to be known as the *smooth model* of ambiguity because, under mild assumptions,

V is a smooth functional, whereas the Choquet expected utility and the maxmin expected utility functionals are typically not everywhere differentiable (over the space of acts).

The notion of second order probabilities is rather old and deserves a separate survey.[40] This idea is at the heart of Bayesian statistics, where Bayes's rule is retained and a probability over probabilities over a state space is equivalent to a probability over the same space. Within decision theory, Segal (1987) already suggested that Ellsberg's paradox can be explained by second-order probabilities, provided that we allow the decision maker to violate the principle of reduction of compound lotteries. Specifically, Segal's model assumed that the second-order probabilities are used to aggregate first-order expectations via Quiggin's (1982) anticipated utility. Other related models have been proposed by Nau (2001, 2006, 2011), Chew and Sagi (2008), Ergin and Gul (2009), and Seo (2009). Halevy and Feltkamp (2005) proposed another approach according to which the decision maker does not err in the computation of probabilities, but uses a mis-specified model, treating a one-shot choice as if it were repeated.

As compared to Choquet expected utility maximization, the smooth preferences model, like the maxmin model, has the advantage of having a simple and intelligible cognitive interpretation. As opposed to both Choquet and maxmin expected utility models, smooth preferences have the disadvantage of imposing non-trivial epistemological demands on the DM: the smooth model requires the specification of a prior over probability models, that is, of a probability μ over a much larger space, $\Delta(\Sigma)$, something that may be informationally and observationally demanding.

That said, beyond the above mentioned separation, the smooth preferences model enjoys an additional advantage of tractability. If S is finite, one may choose μ to be a uniform prior over $\Delta(\Sigma)$ and specify a simple functional form for φ, to get a simple model in which uncertainty/ambiguity attitudes can be analyzed in a way that parallels the treatment of risk attitudes in the classical literature. Specifically, assume that

$$\varphi(x) = -\frac{1}{\alpha}e^{-\alpha x}$$

for $\alpha > 0$. In this case, the DM can be said to have a constant ambiguity aversion α; when $\alpha \to 0$, the DM's preferences converge to Bayesian preferences with prior $\hat{p}$, whereas when $\alpha \to \infty$, preferences converge to MMEU preferences relative to the support of μ. (See Klibanoff et al. 2005, for details.) Thus, the smooth ambiguity aversion model can be viewed as an extension of the maxmin model, in its Waldean interpretation.

[40] Bayes (1763) himself writes in his Proposition 10 that "the chance that the probability of the event lies somewhere between ..." (at the beginning of his essay, in Definition 6 Bayes says that "By chance I mean the same as probability").

Variational Preferences

Maccheroni et al. (MMR, 2006a) suggested and axiomatized an extension of the maxmin model in order to better understand the theoretical foundations of the works of Hansen and Sargent on model uncertainty in macroeconomics (see the surveys Hansen 2007; Hansen and Sargent 2008). These works consider agents who take into account the possibility that their (probabilistic) model Q may not be the correct one, but only an approximation thereof. For this reason, they rank acts f according to the following choice criterion

$$V(f) = \min_{P \in \Delta(\Sigma)} \left\{ \int_S u(f(s)) \, dP(s) + \theta R(P \| Q) \right\}, \tag{21.15}$$

where $\theta > 0$, and $R(\cdot \| Q) : \Delta(\Sigma) \to [0, \infty]$ is the relative entropy with respect to Q.

Preferences $\succsim$ on $\mathcal{F}$ represented by criterion (21.15) are called *multiplier preferences* by Hansen and Sargent. The relative entropy $R(P \| Q)$ measures the relative likelihood of the alternative models P with respect to the reference model Q. The positive parameter θ reflects the weight that agents are giving to the possibility that Q might not be the correct model (as θ becomes larger, agents focus more on Q as the correct model, giving less importance to the alternatives P).

Model uncertainty, which motivated the study of multiplier preferences, is clearly akin to the problem of ambiguity, underlying maxmin preferences. Yet, neither class of preferences is nested in the other. A priori, it was not clear what are the commonalities between these models and how they can be theoretically justified. To address this issue, MMR introduced and axiomatized a novel class of preferences that includes both multiplier and maxmin preferences as special cases.

Specifically, observe that the maxmin criterion (21.10) can be written as

$$V(f) = \min_{P \in \Delta(\Sigma)} \left\{ \int_S u(f(s)) \, dP(s) + \delta_C(P) \right\}, \tag{21.16}$$

where $\delta_C : \Delta \to [0, \infty]$ is the indicator function of C given by

$$\delta_C(P) = \begin{cases} 0 & \text{if} P \in C, \\ \infty & \text{otherwise.} \end{cases}$$

Like the relative entropy, the indicator function is a convex function defined on the simplex $\Delta(\Sigma)$. This suggests the following general representation

$$V(f) = \min_{P \in \Delta(\Sigma)} \left\{ \int u(f(s)) \, dP(s) + c(P) \right\}, \tag{21.17}$$

where $c : \Delta(\Sigma) \to [0, \infty]$ is a convex function on the simplex. MMR call *variational* the preferences $\succsim$ on $\mathcal{F}$ represented by (21.17). Multiplier and maxmin preferences are the special cases of variational preferences where c is, respectively, the relative entropy $\theta R(\cdot \| Q)$ and the indicator function δ_C.

MMR establish a behavioral foundation for the representation (21.17), which in turn offers a common behavioral foundation for multiplier and maxmin preferences. Their axiomatization is based on a relaxation of the C-Independence GS.3 of Gilboa and Schmeidler. To understand it, consider the following equivalent form of GS.3 (C-Independence).

Lemma 13. *A binary relation $\succsim$ on $\mathcal{F}$ satisfies axiom GS.3 (C-Independence) if and only if, for all $f, g \in \mathcal{F}$, $p, q \in \Delta(X)$, and $\alpha, \beta \in (0, 1]$, we have:*

$$\alpha f + (1-\alpha)p \succsim \alpha g + (1-\alpha)p \Rightarrow \beta f + (1-\beta)q \succsim \beta g + (1-\beta)q.$$

Lemma 13 (MMR p. 1454) shows that axiom GS.3 actually involves two types of independence: independence relative to mixing with constants and independence relative to the weights used in such mixing. The next axiom, due to MMR, retains the first form of independence, but not the second one.

MMR.3 WEAK C-INDEPENDENCE: If $f, g \in \mathcal{F}$, $p, q \in \Delta(X)$, and $\alpha \in (0, 1)$,

$$\alpha f + (1-\alpha)p \succsim \alpha g + (1-\alpha)p \Rightarrow \alpha f + (1-\alpha)q \succsim \alpha g + (1-\alpha)q.$$

Axiom MMR.3 is therefore the special case of axiom GS.3 (C-Independence) in which the mixing coefficients α and β are required to be equal. In other words, axiom MMR.3 is a very weak independence axiom that requires independence only with respect to mixing with lottery acts, provided the mixing weights are kept constant.

This is a significant weakening of axiom GS.3 (C-Independence). One might wonder, why would the DM follow MMR.3 but not GS.3 in its full strength. To see this, consider the re-statement of axiom GS.3 in Lemma 13 in the case that the weights α and β are very different, say α is close to 1 and β is close to 0. Intuitively, acts $\alpha f + (1-\alpha)p$ and $\alpha g + (1-\alpha)p$ can then involve far more uncertainty than acts $\beta f + (1-\beta)q$ and $\beta g + (1-\beta)q$, which are almost constant acts. As a result, we expect that, at least in some situations, the ranking between the genuinely uncertain acts $\alpha f + (1-\alpha)p$ and $\alpha g + (1-\alpha)p$ can well differ from that between the almost constant acts $\beta f + (1-\beta)q$ and $\beta g + (1-\beta)q$. By contrast, Axiom MMR.3 is not susceptible to this critique: since only the same coefficient α is used in both sides of the implication, the axiom does not involve acts that differ in their overall uncertainty, as it were.

The representation result of MMR is especially sharp when the utility function u is unbounded (above or below), that is, when its image $u(X) = \{u(x) : x \in X\}$ is an unbounded set. In an AA setup this follows from the following assumption (see Kopylov 2001).

AA.7 UNBOUDEDNESS: There exist $x \succ y$ in X such that for all $\alpha \in (0,1)$ there exists $z \in X$ satisfying either $y \succ \alpha z + (1-\alpha)\, x$ or $\alpha z + (1-\alpha)\, y \succ x$.

We can now state the representation result of MMR, which generalizes Theorem 4 by allowing for general functions $c : \Delta(\Sigma) \to [0, \infty]$. Here x_f denotes the certainty equivalent of act f; i.e., $f \sim x_f$.

Theorem 14. *Let $\succsim$ be a binary relation on $\mathcal{F}$. The following conditions are equivalent:*

(i) $\succsim$ satisfies conditions AA.1, AA.2, MMR.3, AA.4, AA.5, S.6, and AA.7;

(ii) there exists an affine function $u : X \to \mathbb{R}$, with $u(X)$ unbounded, and a grounded,[41] convex, and lower semicontinuous function $c : \Delta(\Sigma) \to [0, \infty]$ such that, for all $f, g \in \mathcal{F}$

$$f \succsim g \Leftrightarrow \min_{P \in \Delta(\Sigma)} \left(\int_S u(f(s))\, dP(s) + c(P) \right) \geq \min_{P \in \Delta(\Sigma)} \left(\int u(g(s))\, dP(s) + c(P) \right). \tag{21.18}$$

For each u there is a unique $c : \Delta(\Sigma) \to [0, \infty]$ satisfying (21.18), given by

$$c(P) = \sup_{f \in \mathcal{F}} \left(u(x_f) - \int_S u(f(s))\, dP(s) \right). \tag{21.19}$$

MMR show how the function c can be viewed as an index of ambiguity aversion, as we will discuss later in section "Ambiguity Aversion". Alternatively, they observe that the function c can be interpreted as the cost of an adversarial opponent of selecting the prior P. In any case, formula (21.19) allows to determine the index c from behavioral (e.g., experimental) data in that it only requires to elicit u and the certainty equivalents x_f.

Behaviorally, maxmin preferences are the special class of variational preferences that satisfy the C-Independence axiom GS.3. For multiplier preferences, however, MMR did not provide the behavioral assumption that characterize them among variational preferences. This question left open by MMR was answered by Strzalecki (2010), who found the sought-after behavioral conditions. They turned out to be closely related to some of Savage's axioms. Strzalecki's findings thus completed the integration of multiplier preferences within the framework of choice under ambiguity.

The weakening of C-Independence in MMR.3 has a natural variation in which independence is restricted to a particular lottery act, but not to a particular weight α. Specifically, one may require that, for the worst possible outcome x_* (if such exists),

[41] The function $c : \Delta(\Sigma) \to [0, \infty]$ is grounded if its infimum value is zero.

$$\alpha f + (1-\alpha)x_* \gtrsim \alpha g + (1-\alpha)x_* \Leftrightarrow \beta f + (1-\beta)x_* \gtrsim \beta g + (1-\beta)x_*$$

for every two acts $f, g \in \mathcal{F}$ and every $\alpha, \beta \in (0, 1]$,

This condition has been used by Chateauneuf and Faro (2009), alongside other conditions, to derive the following representation: there exists a so-called *confidence* function φ on $\Delta(\Sigma)$, and a confidence threshold α, such that acts are evaluated according to

$$V(f) = \min_{\{P \in \Delta(\Sigma) | \varphi(P) \geq \alpha\}} \left[\frac{1}{\varphi(P)} \int_S u(f(s)) dP(s) \right]$$

This decision rule suggests that the DM has a degree of confidence $\varphi(P)$ in each possible prior P. The expected utility associated with a prior P is multiplied by the inverse of the confidence in P, so that a low confidence level is less likely to determine the minimum confidence-weighted expected utility of f.

The intersection of the classes of variational preferences with confidence preferences is the maxmin model, satisfying C-Independence in its full force.[42] See also Ghirardato et al. (2005) for other characterizations of C-Independence.

Beyond Independence: Uncertainty Averse Preferences

All the choice models that we reviewed so far feature some violation of the Independence axiom AA.3, which is the main behavioral assumption questioned in the literature on choice under ambiguity in a AA setup. In order to better understand this class of models, Cerreia-Vioglio et al. (2011) recently established a common representation that unifies and classifies them. Since a notion of minimal independence among uncertain acts is, at best, elusive both at a theoretical and empirical level, this common representation does not use any independence condition on uncertain acts, however weak it may appear.

Cerreia-Vioglio et al. (2011) thus studied uncertainty averse preferences, that is, complete and transitive preferences that are monotone and convex, without any independence requirement on uncertain acts. This general class of preferences includes as special cases variational preferences, confidence preferences, as well as smooth preferences with a concave φ.

Though no independence assumption is made on uncertain acts, to calibrate risk preferences Cerreia-Vioglio et al. assumed standard independence on lottery acts.

CMMM.3 RISK INDEPENDENCE: If $p, q, r \in \Delta(X)$ and $\alpha \in (0, 1)$, $p \sim q \Rightarrow \alpha p + (1-\alpha) r \sim \alpha q + (1-\alpha) r$.

[42] This is so because one axiom relates preferences between mixtures with different coefficients α, β and the other – between mixtures with different constant acts x_*, p.

Along with the other axioms, CMMM.3 implies that the risk preference $\succsim_\Delta$ satisfies the von Neumann-Morgenstern axioms. In the representation result of Cerreia-Vioglio et al. (2011) functions of the form $G : \mathbb{R}\times\Delta(\Sigma) \to (-\infty,\infty]$ play a key role. Denote by $\mathcal{G}(\mathbb{R}\times\Delta(\Sigma))$ the class of these functions such that:

(i) G is quasiconvex on $\mathbb{R}\times\Delta(\Sigma)$,
(ii) $G(\cdot, P)$ is increasing for all $P \in \Delta(\Sigma)$,
(iii) $\inf_{P\in\Delta(\Sigma)} G(t,P) = t$ for all $t \in T$.

We can now state a version of their main representation theorem.

Theorem 15. *Let $\succsim$ be a binary relation on $\mathcal{F}$. The following conditions are equivalent:*

(i) $\succsim$ satisfies axioms AA.1, AA.2, CMMM.3, AA.4, AA.5, S.6, AA.7;
(ii) there exists a non-constant affine $u : X \to \mathbb{R}$, with $u(X) = \mathbb{R}$, and a lower semicontinuous $G : \mathbb{R}\times\Delta(\Sigma) \to (-\infty,\infty]$ that belongs to $\mathcal{G}(\mathbb{R}\times\Delta(\Sigma))$ such that, for all f and g in $\mathcal{F}$,

$$f \succsim g \Leftrightarrow \min_{P\in\Delta(\Sigma)} G\left(\int u(f)\,dP, P\right) \geq \min_{P\in\Delta(\Sigma)} G\left(\int u(g)\,dP, P\right). \quad (21.20)$$

The function u is cardinally unique and, given u, the function G in (21.20) is given by

$$G(t,P) = \sup_{f\in\mathcal{F}} \left\{ u\left(x_f\right) : \int u(f)\,dP \leq t \right\} \qquad \forall\,(t,p) \in \mathbb{R}\times\Delta(\Sigma). \quad (21.21)$$

In this representation DMs can be viewed as if they considered, through the term $G\left(\int u(f)\,dP, P\right)$, all possible probabilities P and the associated expected utilities $\int u(f)\,dP$ of act f. They then behave as if they summarized all these evaluations by taking their minimum. The quasiconvexity of G and the cautious attitude reflected by the minimum in (21.20) derive from the convexity of preferences. Their monotonicity, instead, is reflected by the monotonicity of G in its first argument.

The representation (21.20) features both probabilities and expected utilities, even though no independence assumption whatsoever is made on uncertain acts. In other words, this representation establishes a general connection between the language of preferences and the language of probabilities and utilities, in keeping with the tradition of the representation theorems in choice under uncertainty.

Cerreia-Vioglio et al. (2011) show that G can be interpreted as index of uncertainty aversion, in the sense of section "Ambiguity Aversion" below. Moreover, (21.21) shows that this index can be elicited from choice behavior.

Variational preferences correspond to additively separable functions G, i.e., these preferences are characterized by

$$G(t,P) = t + c(P)$$

where $c : \Delta(\Sigma) \to [0, \infty]$ is a convex function. In this case (21.20) reduces to the variational representation (21.18).

Smooth preferences with concave ϕ correspond to the uncertainty aversion index given by

$$G(t, P) = t + \min_{\nu \in \Gamma(P)} I_t(\nu \parallel \mu) \tag{21.22}$$

where $I_t(\cdot \parallel \mu)$ is a suitable statistical distance function that generalizes the classic relative entropy, and $\Gamma(P)$ is the set of all second-order probabilities ν that are absolutely continuous with respect to μ and that have P as their reduced, first-order, probability measure on S.

Other Classes of Preferences

The scope of this paper does not allow us to do justice to the variety of decision models that have been suggested in the literature to deal with uncertainty in a non-probabilistic way, let alone the otherwise growing literature in decision theory.[43] Here we only mention a few additional approaches to the problem of ambiguity.

As mentioned above, Segal (1987, 1990) suggested a risk-based approach to uncertainty, founded on the idea that people do not reduce compound lotteries. Recently, Halevy (2007) provided some experimental evidence on the link between lack of reduction of compound lotteries and ambiguity, and Seo (2009) carried out an in depth theoretical analysis of this issue. Since failure to reduce compound lotteries is often regarded as a mistake, this source of ambiguity has a stronger positive flavor than the absence of information, which is our main focus.

Stinchcombe (2003), Olszewski (2007), and Ahn (2008) model ambiguity through sets of lotteries, capturing exogenous or objective ambiguity. (See also Jaffray (1988), who suggested related ideas). Preferences are defined over these sets, with singleton and nonsingleton ones modelling risky and ambiguous alternatives, respectively. For example, these sets can be ranked either according to the criterion $V(A) = \left(\int_A \phi \circ u d\mu\right) / \mu(A)$ where ϕ and μ model ambiguity attitudes (Ahn 2008) or the criterion $V(A) = \alpha \min_{l \in A} U(l) + (1 - \alpha) \max_{l \in A} U(l)$ where α models ambiguity attitudes (Olszewski 2007). Viero (2009) combines this approach with the Anscombe-Aumann model.

Chateauneuf et al. (2007) axiomatize *neo-additive Choquet expected utility*, a tractable CEU criterion of the "Hurwicz" form $V(f) = \alpha \int u(f(s)) dP(s) + \beta \max_s u(f(s)) + (1 - \alpha - \beta) \min_s u(f(s))$. Through the values of the weights α and β, the preference functional V captures in a simple way different degrees of optimism and pessimism, whose extreme forms are given by the min and max of $u(f(s))$.

[43] Other sub-fields include choices from menus, decision under risk, minmax regret approaches, and others. On the first of these, see Limpan and Pesendorfer (2013).

Gajdos et al. (2008) axiomatize, as discussed before, a model with objective information. Preferences are defined over pairs (f, C) of acts and sets of probabilities (that represent objective information). Such pairs are ranked through the functional $V(f, C) = \min_{p \in \varphi(C)} \int u(f(s))\, dP(s)$, where $\varphi(C) \subseteq C$ is the subset of C that we denoted in the earlier discussion as D. When $\varphi(C) = C$, we get back to the MMEU model.

Gul and Pesendorfer (2008) suggested *subjective expected uncertain utility theory*, according to which acts can be reduced to bilotteries, each specifying probabilities for ranges of outcome values, where these probabilities need not be allocated to sub-ranges. Arlo-Costa and Helzner (2010a) propose to deal with the comparative ignorance hypothesis of Tversky and Fox (1995), and present experimental findings that challenge the explanation provided by the latter. (See also Arlo-Costa and Helzner 2010b).

Siniscalchi (2009a) axiomatizes *vector expected utility*, in which Savage's acts are assessed according to $V(f) = \int u(f(s))\, dP(s) + A\left(\left(\int \xi_i \cdot u(f(s))\, dP(s)\right)_{i=1,\ldots,n}\right)$ where the first term on the right hand side is a baseline expected-utility evaluation and the second term is an adjustment that reflects DMs' perception of ambiguity and their attitudes toward it. In particular, ξ_i are random variables with zero mean that model different sources of ambiguity (see Siniscalchi 2009a, p. 803).

Given the variety of the models of decision making that allow for non-neutral approaches to ambiguity, one is led to ask, how should we select a model to work with? There are at least three possible approaches to this problem. First, one may follow the classical empirical tradition and compare the different models by a "horse-race" . The model that best explains observed phenomena should be used for prediction, with the usual trade-offs between the model's goodness of fit and its simplicity and generality. The degree to which models fit the data should be measured both for their assumptions and for their conclusions. (Indeed, the assumptions are also, in a trivial sense, conclusions.) Thus, this approach calls both for experimental tests of particular axioms and of entire models, as well as for empirical tests of theories based on these models. Importantly, when engaging in such an endeavor, one should be prepared to find that a model may be the most appropriate for analyzing certain phenomena but not for others. Thus, for example, it is possible that smooth preferences are the best model for the behavior of organizations, whereas variational preferences are a better description for the behavior of individuals. Or that labor search models are best explained by the maxmin model, while financial investments call for the Hurwicz-Jaffray model, and so forth.

For qualitative analysis, one may adopt a second approach, which does not commit to a particular model of decision under uncertainty, but uses representatives of these models in order to gain robust insights. Adopting this approach, a researcher may start with a benchmark Bayesian model, and add the uncertainty ingredient using any of the models mentioned above, as a sensitivity analysis of the Bayesian model. In this approach, theoretical convenience may be an important guideline. However, it will be advisable to trust only the qualitative conclusions that emerge from more than one model. That is, sensitivity analysis itself should not be too sensitive.

Finally, in light of the variety of models and the theoretical difficulties in selecting a single one, one may choose a third approach, which attempts to obtain general conclusions within a formal model, without committing to a particular theory of decision making. This approach has been suggested in the context of risk by Machina (1982). In this celebrated paper, facing a variety of decision models under risk, Machina attempted to show that much of economic analysis of choice under risk can be carried through without specifying a particular model. More concretely, Machina stipulated a functional on lotteries (with given probabilities) that was smooth enough to allow local approximations by linear functions. The gradient of the functional was considered to be a *local* utility function. Machina has shown that some results in economic theory could be derived by allowing the local utility function to vary, as long as it satisfied the relevant assumptions. Machina's approach was therefore not about decision theory per se. It was about the degree to which decision theory mattered: it showed that, for some applications, economists need not worry about how people really make decisions, since a wide range of models were compatible with particular qualitative conclusions.

A similar approach has been suggested for decisions under uncertainty. An early example of this approach is the notion of biseparable preferences, suggested by Ghirardato and Marinacci (2001), and mentioned above. *Biseparable preferences* are any monotone and continuous preferences over general acts that, when restricted to acts f with only two outcomes, say, x and y, can be described by the maximization of

$$J(f) = u(x)\nu(A) + (u(x) - u(y))(1 - \nu(A))$$

where ν is a capacity and

$$f(s) = \begin{cases} x & s \in A \\ y & s \notin A \end{cases}$$

with $x \succ y$. Biseparable preferences include both CEU and MMEU. Ghirardato and Marinacci (2001) provide a definition of uncertainty aversion that does not depend on the specific model of decision making and applies to all biseparable preferences.

More recently, Machina (2005) suggested a general approach to preferences under uncertainty which, similarly to Machina (1982), assumes mostly smoothness and monotonicity of preferences, but remains silent regarding the actual structure of preferences, thereby offering a highly flexible model.

Ambiguity Aversion

Schmeidler's axiom S.6 provided a first important characterization of ambiguity aversion, modelled through a preference for hedging/randomization. Epstein (1999) and Ghirardato and Marinacci (2002) studied this issue from a different perspective, inspired by Yaari (1969)'s analysis of comparative risk attitudes.

Here we present the approach of Ghirardato and Marinacci because of its sharper model implications. This approach relies on two key ingredients:

(i) A comparative notion of ambiguity aversion that, given any two preferences $\succsim_1$ and $\succsim_2$ on $\mathcal{F}$, says when $\succsim_1$ is more ambiguity averse than $\succsim_2$.
(ii) A benchmark for neutrality to ambiguity; that is, a class of preferences $\succsim$ on $\mathcal{F}$ that are viewed as neutral to ambiguity.

The choice of these ingredients in turn determines the absolute notion of ambiguity aversion, because a preference $\succsim$ on $\mathcal{F}$ is classified as ambiguity averse provided it is more ambiguity averse than an ambiguity neutral one.

The comparative notion (i) is based on comparisons of acts with lottery acts that deliver a lottery p at all states. We consider them here because they are the most obvious example of unambiguous acts, that is, acts whose outcomes are not affected by the unknown probabilities.

Consider DM_1 and DM_2, whose preferences on $\mathcal{F}$ are $\succsim_1$ and $\succsim_2$, respectively. Suppose that

$$f \succsim_1 p,$$

that is, DM_1 prefers the possibly ambiguous act f to the unambiguous lottery act p. If DM_1 is more ambiguity averse than DM_2 it is natural to expect that DM_2 will also exhibit such preferences:

$$f \succsim_2 p.$$

For, if DM_1 is bold enough to have $f \succsim_1 p$, then DM_2 – who dislikes ambiguity no more than DM_1 – must be at least equally bold.

We take this as the behavioral characterization of the comparative notion of ambiguity aversion.

Definition 16. Given two preferences $\succsim_1$ and $\succsim_2$ on $\mathcal{F}$, $\succsim_1$ is *more ambiguity averse than* $\succsim_2$ if, for all $f \in \mathcal{F}$ and $p \in \Delta(X)$,

$$f \succsim_1 p \Rightarrow f \succsim_2 p. \tag{21.23}$$

As benchmark for neutrality to ambiguity we consider subjective expected utility (SEU) preferences on $\mathcal{F}$. These preferences intuitively embody ambiguity neutrality. They might not be the only preference embodying ambiguity neutrality, but they seem to be the most obvious ones.[44]

Methodologically, like the choice of lottery acts as the unambiguous acts in the comparison (21.23), also the neutrality benchmark is chosen by making the weakest

[44] Epstein (1999) takes the standard for ambiguity neutrality to be preferences that are probabilistically sophisticated in the sense of Machina and Schmeidler (1992). In his approach Theorem 18 below does not hold.

prejudgment on which preferences qualify for this role. Sharp model implications will follow, nevertheless, as we will see momentarily.

Having thus prepared the ground, we can define ambiguity aversion

Definition 17. A preference relation $\succsim$ on $\mathcal{F}$ is *ambiguity averse* if it is more ambiguity averse than some SEU preference on $\mathcal{F}$.

The next result, due to Ghirardato and Marinacci (2002), applies these notions to the maxmin expected utility (MEU) model. Here $u_1 \approx u_2$ means that there exist $\alpha > 0$ and $\beta \in \mathbb{R}$ such that $u_1 = \alpha u_2 + \beta$.

Theorem 18. *Given any two MMEU preferences $\succsim_1$ and $\succsim_2$ on $\mathcal{F}$, the following conditions are equivalent:*

(i) $\succsim_1$ *is more ambiguity averse than* $\succsim_2$,
(ii) $u_1 \approx u_2$ *and* $C_1 \subseteq C_2$ *(provided* $u_1 = u_2$*).*

Given that $u_1 \approx u_2$, the assumption $u_1 = u_2$ is just a common normalization of the two utility indices. Therefore, Theorem 18 says that more ambiguity averse MMEU preferences are characterized, up to a normalization, by smaller sets of priors C. Therefore, the set C can be interpreted as an *index of ambiguity aversion*.

This result thus provides a behavioral foundation for the comparative statics exercises in ambiguity through the size of the sets of priors C that play a key role in the economic applications of the MMEU model. In fact, a central question in these applications is how changes in ambiguity attitudes affect the relevant economic variables.

An immediate consequence of Theorem 18 is that, not surprisingly, MMEU preferences are always ambiguity averse. That is, they automatically embody a negative attitude toward ambiguity, an attitude inherited from axiom S.6.

The condition $u_1 \approx u_2$ ensures that risk attitudes are factored out in comparing the MMEU preferences $\succsim_1$ and $\succsim_2$. This is a dividend of the risk calibration provided by the AA setup via the risk preference $\succsim_\Delta$ discussed in section "The Anscombe-Aumann Setup". In a Savage setup, where this risk calibration is no longer available, Definition 16 has to be enriched in order to properly factor out risk attitudes, so that they do not interfere with the comparison of ambiguity attitudes (see Ghirardato and Marinacci (2002), for details on this delicate conceptual issue).

Maccheroni et al. (2006a) generalize Theorem 18 to variational preferences by showing that the condition $C_1 \subseteq C_2$ takes in this case the more general form $c_1 \leq c_2$. The function c can thus be viewed as an index of ambiguity aversion that generalizes the sets of priors C. Variational preferences are always ambiguity averse, a fact that comes as no surprise since they satisfy axiom S.6.

For CEU preferences, Ghirardato and Marinacci (2002) show that more ambiguity averse CEU preferences are characterized, up to a common normalization of utility indexes, by smaller capacities ν. More interestingly, they show that CEU preferences are ambiguity averse when the cores of the associated capacities are nonempty. Since convex capacities have nonempty cores, CEU preferences that satisfy axiom S.6 are thus ambiguity averse. The converse, however, is not

true since there are capacities with nonempty cores that are not convex. Hence, there exist ambiguity averse CEU preferences that do not satisfy S.6, which is thus a sufficient but not necessary condition for the ambiguity aversion of CEU preferences. Ghirardato and Marinacci (2002) discuss at length this feature of CEU preferences, and we refer the interested reader to that paper for details (see also Chateauneuf and Tallon (2002), who present a notion of weak ambiguity aversion for CEU preferences, as well as Montesano and Giovannone (1996), who investigate how CEU preferences may reflect aversion to increasing ambiguity).

Unambiguous events Unambiguous events should be events over which decision makers do not perceive any ambiguity. Intuitively, in terms of functional forms an event E is unambiguous for a preference $\succsim$ if:

(i) $\nu(E) + \nu(E^c) = 1$ when $\succsim$ is CEU;
(ii) $P(E) = P'(E)$ for all $P, P' \in C$ when $\succsim$ is MMEU and, more generally, for all $P, P' \in \operatorname{dom} c$ when $\succsim$ is variational;[45]
(iii) $p(E) = k\mu$-a.e. for some $k \in [0, 1]$ when $\succsim$ is smooth.

A few behavioral underpinnings of these notions of unambiguous event have been proposed by Nehring (1999), Epstein and Zhang (2001), Ghirardato and Marinacci (2002), Zhang (2002), Ghirardato et al. (2004), Klibanoff et al. (2005), and Amarante and Feliz (2007) (who also provide a discussion of some of the earlier notions which we refer the interested to).

Updating Beliefs

How should one update one's beliefs when new information is obtained? In the case of probabilistic beliefs there is an almost complete unanimity that Bayes's rule is the only sensible way to update beliefs. Does it have an equivalent rule for the alternative models discussed above? The answer naturally depends on the particular non-Bayesian model one adopts. At the risk of over-generalizing from a small sample, we suggest that Bayes's rule can typically be extended to non-Bayesian beliefs in more than one way. Since the focus of this survey is on static preferences, we mention only a few examples, which by no means exhaust the richness of dynamic models.

For instance, if one's beliefs are given by a capacity ν, and one learns that an event B has obtained, one may assign to an event A the weight corresponding to the straightforward adaptations of Bayes's formula:

$$\nu(A|B) = \frac{\nu(A \cap B)}{\nu(B)}$$

[45] $\operatorname{dom} c$ is the effective domain of the function c; i.e., $\operatorname{dom} c = \{P \in \Delta(S) : c(p) < +\infty\}$.

However, another formula has been suggested by Dempster (1967, see also Shafer 1976) as a special case of his notion of merging of belief functions:

$$\nu(A|B) = \frac{\nu((A \cap B) \cup B^c) - \nu(B^c)}{1 - \nu(B^c)}$$

Clearly, this formula also boils down to standard Bayesian updating in case ν is additive. Yet, the two formulae are typically not equivalent if the capacity ν fails to be additive. Each of these formulae extends some, but not all, of the interpretations of Bayesian updating from the additive to the non-additive case.

If beliefs are given by a set of priors C, and event B is known to have occurred, a natural candidate for the set of priors on B is simply the same set C, where each probability is updated according to Bayes's rule. This results in *full Bayesian updating (FBU)*, defining the set of priors (on B)

$$C_B = \{p(\cdot|B) \mid p \in C\}$$

FBU allows standard learning given each possible prior, but does not reflect any learning about the set of priors that should indeed be taken into consideration. It captures Bayesian learning (conditional on a prior) but not the statistical inference typical of classical statistics, namely, the selection of subsets of distributions from an a priori given set of distributions. If we were to think of each prior p in C as an expert, who expresses her probabilistic beliefs, FBU can be interpreted as if each expert were learning from the evidence B, while the DM does not use the evidence to decide which experts' advice to heed.[46]

Following this line of reasoning, and in accordance with statistical principles, one may wish to select probabilities from the set C based on the given event B. One, admittedly extreme way of doing so is to adopt the maximum likelihood principle. This suggests that only the priors that a priori used to assign the highest probability to the event B should be retained among the relevant ones. Thus, *maximum likelihood updating (MLU)* is given by

$$C_B^M = \left\{ p(\cdot|B) \;\middle|\; p \in \arg\max_{q \in C} q(B) \right\}$$

If one's beliefs are given by a convex capacity, or, equivalently, by a set C which is the core of a convex capacity, MLU is equivalent to Dempster-Shafer's updating. This rule has been axiomatized by Gilboa and Schmeidler (1993), whereas FBU, suggested by Jean-Yves Jaffray, has been axiomatized by Pires (2002).

FBU and MLU are both extreme. Using the experts metaphor, FBU retains all experts, and gives as much weight to those who were right as to those who were

[46] See Seidenfeld and Wasserman (1993) who study counter-intuitive updating phenomena in this context.

practically proven wrong in their past assessments. By contrast, MLU completely ignores any expert who was not among the maximizers of the likelihood function. It therefore makes sense to consider intermediate methods, though, to the best of our knowledge, none has been axiomatized to date.

The tension between FBU and MLU disappears if the set of priors C is *rectangular* (Epstein and Miao 2003) in the sense that it can be decomposed into a set of current-period beliefs, coupled with next-period conditional beliefs, in such a way that any combination of the former and the latter is in C. Intuitively, rectangularity can be viewed as independence of sorts: it suggests that whatever happens in the present period does not teach us which prior (or expert) is to be trusted more in the next period. Formally, the set of conditional probabilities on the given event B using all priors and the set obtained using only the maximum likelihood ones coincide. Related arguments, in particular how rectangular sets of priors would lead to consistent dynamic MMEU behavior, were made by Sarin and Wakker (1998) (see in particular their Theorem 2.1). See also Epstein and Schneider (2007), who consider updating in a more explicit model, distinguishing between the set of parameters and the likelihood functions they induce.

Epstein and Miao (2003) consider preferences over consumption processes, and axiomatize a decision rule that extends MMEU to the dynamic set-up recursively. Their axioms also guarantee that the set of priors C is rectangular. The recursive structure means that the maxmin expected utility at a given period for the entire future can also be written as maxmin expected utility over the present period and the discounted continuation (MMEU) value starting in the following period. Wang (2003b) proposed a related recursive approach.

This recursive approach extends beyond the MMEU model. It has similarly been applied to extend smooth preferences (see Hayashi and Miao 2011; Klibanoff et al. 2009) and variational preferences to dynamic set-ups (see Maccheroni et al. 2006b). Equipped with a variety of models of behavior with ambiguous beliefs, which are adapted to deal with dynamic problems recursively, the stage is set to analyze economic problems in not-necessarily Bayesian ways.

Another approach to updating was proposed by Hanany and Klibanoff (2007, 2009). They retain dynamic consistency by allowing the update rule to depend not only on original beliefs and new information, but also on the choice problem. In the case of the MMEU model, their approach consists of selecting a subset of priors, and updating them according to Bayes rule, while the relevant subset of priors generally depends on the act chosen before the arrival of new information.

A different route was pursued by Siniscalchi (2006b), who investigated choices over decision trees rather than over temporal acts. This modification allows him to consider sophisticated choices, characterized through a natural notion of consistent planning, under ambiguity.

An important problem relating to updating is the long-run behavior of beliefs. Suppose that a non-Bayesian decision maker faces a process that is, in a well-defined sense, repeated under the same conditions. Will she learn the true process? Will the set of probabilities converge in the limit to the true one? A partial answer was given in the context of capacities, where laws of large numbers have been proved by

Marinacci (1999, 2002b) and Maccheroni and Marinacci (2005). The behavior of the set of probabilities in the context of the maxmin model was analyzed in Epstein and Schneider (2007).

Applications

There are many economic models that lead to different qualitative conclusions when analyzed in a Bayesian way as compared to the alternative, non-Bayesian theories. The past two decades have witnessed a variety of studies that re-visited classical results and showed that they need to be qualified when one takes ambiguity into account. The scope of this paper allows us to mention but a fraction of them. The following is a very sketchy description of a few studies, designed only to give a general idea of the scope of theoretical results that need to be re-examined in light of the limitations of the Bayesian approach.[47]

Dow and Werlang (1992) analyzed a simple asset pricing model. They showed that, if an economic agent is ambiguity averse as in the CEU or MMEU model, then there will be a *range* of prices at which she will wish neither to buy nor to sell a financial asset. This range will be of non-zero length even if one ignores transaction costs. To see the basic logic of this result, consider two states of the world, where the probability of the first state, p, is only known to lie in the interval $[0.4, 0.6]$. (This will also be the core of a convex capacity). Assume that a financial asset X yields 1 in the first state and -1 in the second. The MMEU model values both X and $-X$ at -0.2. In a Bayesian model, p would be known, and the agent would switch, at a certain price π, from demanding X to offering it. This is no longer the case when p is not known. In this case, assuming ambiguity aversion, there will be an interval of prices π at which neither X nor $-X$ will seem attractive to the agent. This may explain why people refrain from trading in certain markets. It can also explain why at times of greater volatility one may find lower volumes of trade: with a larger set of probabilities that are considered possible, there will be more DMs who prefer neither to buy nor to sell.[48] The question of trade among uncertainty averse agents has been also studied in Billot et al. (2000), Kajii and Ui (2006, 2009), and Rigotti et al. (2008).

Epstein and Miao (2003) use uncertainty aversion to explain the home bias phenomenon in international finance, namely, the observation that people prefer to trade stocks of their own country rather than foreign ones. The intuition is that agents know the firms and the stock market in their own country better than in foreign ones. Thus, there is more ambiguity about foreign equities than about domestic

[47] Mukerji and Tallon (2004) survey early works in this area.

[48] This argument assumes that the decision maker starts with a risk-free portfolio. A trader who already holds an uncertain position may be satisfied with it with a small set of probabilities, but wish to trade in order to reduce uncertainty if the set of probabilities is larger.

ones. A Bayesian analysis makes it more difficult to explain this phenomenon: when a Bayesian DM does not know the distribution of the value of a foreign equity, she should have beliefs over it, reducing uncertainty to risk. Thus, a Bayesian would behave in qualitatively similar ways when confronting known and unknown distributions. By contrast, the notion that agents are ambiguity averse may more readily explain why they prefer to trade when the value distribution is closer to being known than when there is a great deal of ambiguity about it.

There are many other applications of ambiguity aversion to models of asset pricing. For example, Epstein and Schneider (2008) show that models involving ambiguity can better capture market reaction to the quality of information than can Bayesian models (see also Epstein and Schneider 2010), while Gollier (2011) shows that ambiguity aversion may not reinforce risk aversion and investigates how this may affect asset prices. Other recent asset pricing applications include Garlappi et al. (2007), Caskey (2009), Miao (2009), Ju and Miao (2012), and Miao and Wang (2011) (see also Guidolin and Rinaldi 2013).

The MMEU model has also been employed in a job search model by Nishimura and Ozaki (2004). They ask how an unemployed agent will react to increasing uncertainty in the labor market. In a Bayesian model, greater uncertainty might be captured by higher variance of the job offers that the agent receives. Other things being equal, an increase in variance should make the agent less willing to accept a given offer, knowing that he has a chance to get better ones later on. This conclusion is a result of the assumption that all uncertainty is quantifiable by a probability measure. Nishimura and Ozaki (2004) show that for an ambiguity averse agent, using the MMEU model, the conclusion might be reversed: in the presence of greater uncertainty, modeled as a larger set of possible priors, the agent will be more willing to take a given job offer rather than bet on waiting for better ones in the future.

Hansen et al. (1999, 2002) compare savings behavior under expected utility maximization with savings behavior of a *robust DM* who behaves in accordance with the multiple prior model. They show that the behavior of a robust DM puts the market price of risk much closer to empirical estimates than does the behavior of the classical expected utility maximizer, and, in particular, can help account for the equity premium. Hansen and Sargent (2001, 2008) apply multiplier preferences to macroeconomic questions starting from the viewpoint that, whatever the probability model a policy maker might have, it cannot be known with certainty. They ask how robust economic policy would be to variations in the underlying probability, and find conclusions that differ qualitatively from classical results. See also Miao (2004), who studies the consumption-savings decision in a different set-up.

Other (published) applications of ambiguity averse preferences include Epstein and Wang (1994, 1995), who explain financial crashes and booms, Mukerji (1998), who explains incompleteness of contracts, Chateauneuf et al. (2000), who study optimal risk-sharing rules with ambiguity averse agents, Greenberg (2000), who finds that in a strategic set-up a player may find it beneficial to generate ambiguity about her strategy choice, Mukerji and Tallon (2001), who show how incompleteness of financial markets my arise because of ambiguity aversion, Rigotti and Shannon (2005), who characterize equilibria and optima and study how they depend

on the degree of ambiguity, Bose et al. (2006), who study auctions under ambiguity, Nishimura and Ozaki (2007), who show that an increase in ambiguity changes the value of an investment opportunity differently than does an increase in risk, Easley and O'Hara (2009, 2010), who study how ambiguity affects market participation, and Treich (2010), who studies when the value of a statistical life increases under ambiguity aversion.

As mentioned above, this list is but a sample of applications and has no claim even to be a representative sample.

Conclusion

Uncertainty is present in practically every field of economic enquiry. Problems in growth and finance, labor and development, political economy and industrial organization lead to questions of uncertainty and require its modeling.

For the most part, economic theory has strived to have a unifying approach to decision making in general, and to decision under uncertainty in particular. It is always desirable to have simple, unifying principles, especially if, as is the case with expected utility theory, these principles are elegant and tractable.

At the same time, expected utility theory appears to be too simple for some applications. Despite its considerable generality, there are phenomena that are hard to accommodate with the classical theory. Worse still, using the classical theory alone may lead to wrong qualitative conclusions, and may make it hard to perceive certain patterns of economic behavior that may be readily perceived given the right language.

At this point it is not clear whether a single paradigm of decision making under uncertainty will ever be able to replace the Bayesian one. It is possible that different models will prove useful to varying degrees in different types of problems. But even if a single paradigm will eventually emerge, it is probably too soon to tell which one it will be.

For the time being, it appears that economic theory may benefit from having more than a single theory of decision under uncertainty in its toolbox. The Bayesian model is surely a great candidate to remain the benchmark. Moreover, often it is quite obvious that the insights learned from the Bayesian analysis suffice. For example, Akerlof's (1970) lemons model need not be generalized to incorporate ambiguity. Its insight is simple and clear, and it will survive in any reasonable model. But there are other models in which the Bayesian analysis might be misleadingly simple. In some cases, adding a touch of ambiguity to the model, often in whatever model of ambiguity one fancies, suffices to change the qualitative conclusions. Hence it seems advisable to have models of ambiguous beliefs in our toolbox, and to test each result, obtained under the Bayesian assumptions, for robustness relative to the presence of ambiguity.

Acknowledgements We thank Giulia Brancaccio, Simone Cerreia-Vioglio, Fabio Maccheroni, Andrew Postlewaite, Xiangyu Qu, and David Schmeidler for comments on earlier drafts of this survey. We are also grateful to many members of the "decision theory forum" for additional comments and references. Finally, we are indebted to Eddie Dekel for many comments and suggestions. Gilboa gratefully acknowledges the financial support of the Israel Science Foundation (grant 396/10) and of the European Reseach Council (advanced grant 269754), and Marinacci that of the European Reseach Council (advanced grant BRSCDP-TEA).

References

Ahn, D. (2008). Ambiguity without a state space. *Review of Economic Studies, 75*, 3–28.

Akerlof, G. A. (1970). The market for 'Lemons': Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics, 84*, 488–500.

Al-Najjar, N., & Weinstein, J. L. (2009). The ambiguity aversion literature: A critical assessment. *Economics and Philosophy, 25*, 249–284.

Alon, S., & Schmeidler, D. (2014). Purely subjective maxmin expected utility. *Journal of Economic Theory, 152,* 382–412

Amarante, M. (2009). Foundations of Neo-Bayesian statistics. *Journal of Economic Theory, 144*, 2146–2173.

Amarante, M., & Feliz, E. (2007). Ambiguous events and maxmin expected utility. *Journal of Economic Theory, 134*, 1–33.

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of Mathematics and Statistics, 34*, 199–205.

Arlo-Costa, H., & Helzner, J. (2010a). Ambiguity aversion: The explanatory power of indeterminate probabilities. *Synthese, 172*, 37–55.

Arlo-Costa, H., & Helzner, J. (2010b). Ellsberg choices: Behavioral anomalies or new normative insights? *Philosophy of Science, 3*, 230–253.

Arrow, K. J. (1970). *Essays in the theory of risk-bearing*. Amsterdam: North-Holland.

Arrow, K. J., & Hurwicz, L. (1972). An optimality criterion for decision making under ignorance. In C. F. Carter & J. L. Ford (Eds.), *Uncertainty and expectations in economics*. Oxford: Basil Blackwell.

Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1997). Thinking coherently. *Risk, 10*, 68–71.

Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance, 9*, 203–228.

Aumann, R. J. (1962). Utility theory without the completeness axiom. *Econometrica, 30*, 445–462.

Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics, 1*, 67–96.

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics, 4*, 1236–1239.

Aumann, R. J. (1987). Correlated equilibrium as an expression of bayesian rationality. *Econometrica, 55*, 1–18.

Bayarri, M. J., & Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science, 19*, 58–80.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370–418.

Berger, J. (2004). The case for objective bayesian analysis. *Bayesian Analysis, 1*, 1–17.

Bernoulli, J. (1713). *Ars Conjectandi*. Basel: Thurneysen Brothers (trans. E. D. Sylla, *The art of conjecturing*. Johns Hopkins University Press, 2005).

Bertrand, J. (1907). *Calcul de probabilite* (2nd ed.). Paris: Gauthiers Villars.

Bewley, T. (2002). Knightian decision theory: Part I. *Decisions in Economics and Finance, 25*, 79–110. (Working paper, 1986).

Billot, A., Chateauneuf, A., Gilboa, I., & Tallon, J.-M. (2000). Sharing beliefs: Between agreeing and disagreeing. *Econometrica, 68*, 685–694.

Bose, S., Ozdenoren, E., & Pape, A. (2006). Optimal auctions with ambiguity. *Theoretical Economics, 1*, 411–438.

Bridgman, P. W. (1927). *The logic of modern physics*. New York: Macmillan.

Carnap, R. (1923). Uber die Aufgabe der Physik und die Andwednung des Grundsatze der Einfachstheit. *Kant-Studien, 28*, 90–107.

Casadesus-Masanell, R., Klibanoff, P., & Ozdenoren, E. (2000). Maxmin expected utility over savage acts with a set of priors. *Journal of Economic Theory, 92*, 35–65.

Caskey, J. (2009). Information in equity markets with ambiguity-averse investors. *Review of Financial Studies, 22*, 3595–3627.

Castagnoli, E., Maccheroni, F., Marinacci, M. (2003). Expected utility with multiple priors. In *Proceedings of ISIPTA 2003*, Lugano.

Cerreia-Vioglio, S., Maccheroni, F., Marinacci, M., & Montrucchio, L. (2011). Uncertainty averse preferences. *Journal of Economic Theory, 146*(4), 1275–1330.

Chateauneuf, A., Dana, R.-A., & Tallon, J.-M. (2000). Optimal risk-sharing rules and equilibria with Choquet expected utility. *Journal of Mathematical Economics, 34*, 191–214.

Chateauneuf, A., Eichberger, J., & Grant, S. (2007). Choice under uncertainty with the best and worst in mind: Neo-additive capacities. *Journal of Economic Theory, 137*, 538–567.

Chateauneuf, A., Maccheroni, F., Marinacci, M., & Tallon, J.-M. (2005). Monotone continuous multiple priors. *Economic Theory, 26*, 973–982.

Chateauneuf, A., & Faro, J. H. (2009). Ambiguity through confidence functions. *Journal of Mathematical Economics, 45*, 535–558.

Chateauneuf, A., & Tallon, J.-M. (2002). Diversification, convex preferences, and non-empty core in the Choquet expected utility model. *Economic Theory, 19*, 509–523.

Chew, H. S., & Sagi, J. (2008). Small worlds: Modeling attitudes toward sources of uncertainty. *Journal of Economic Theory, 139*, 1–24.

Choquet, G. (1953). Theory of capacities. *Annales de l'Institut Fourier, 5*, 131–295.

Cifarelli, D. M., & Regazzini, E. (1996). de Finetti's contribution to probability and statistics. *Statistical Science, 11*, 253–282.

Cyert, R. M., & DeGroot, M. H. (1974). Rational expectations and bayesian analysis. *Journal of Political Economy, 82*, 521–536.

Daston, L. (1995). *Classical probability in the enlightenment*. Princeton: Princeton University Press.

de Finetti, B. (1931). Sul Significato Soggettivo della Probabilità. *Fundamenta Mathematicae, 17*, 298–329.

de Finetti, B. (1937). La Prevision: ses Lois Logiques, ses Sources Subjectives. *Annales de l'Institut Henri Poincare, 7*, 1–68. (trans. H. E. Kyburg & H. E. Smokler (Eds.) *Studies in subjective probability*. Wiley, 1963).

DeGroot, M. H. (1975). *Probability and statistics*. Reading: Addison-Wesley.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics, 38*, 325–339.

Denneberg, D. (1994). *Non-additive measure and integral*. Dordrecht: Kluwer.

Dow, J., & Werlang, S. R. C. (1992). Uncertainty aversion, risk aversion, and the optimal choice of portfolio. *Econometrica, 60*, 197–204.

Easley, D., & O'Hara, M. (2009). Ambiguity and nonparticipation: The role of regulation. *Review of Financial Studies, 22*, 1817–1843.

Easley, D., & O'Hara, M. (2010). Microstructure and ambiguity. *Journal of Finance, 65*, 1817–1846.

Eichberger, J., Grant, S., & Kelsey, D. (2008). Differentiating ambiguity: An expository note. *Economic Theory, 36*, 327–336.

Eichberger, J., Grant, S., Kelsey, D., & Koshevoy, G. A. (2011). The α-MEU model: A comment. *Journal of Economic Theory, 146*(4), 1684–1698.

Efron, B. (1986). Why Isn't everyone a Bayesian? *The American Statistician, 40*, 1–11. With discussion.
Ellsberg, D. (1961). Risk, ambiguity and the savage axioms. *Quarterly Journal of Economics, 75*, 643–669.
Epstein, L. (1999). A definition of uncertainty aversion. *Review of Economic Studies, 66*, 579–608.
Epstein, L. G., & Marinacci, M. (2007). Mutual absolute continuity of multiple priors. *Journal of Economic Theory, 137*, 716–720.
Epstein, L. G., Marinacci, M., & Seo, K. (2007). Coarse contingencies and ambiguity. *Theoretical Economics, 2*, 355–394.
Epstein, L. G., & Miao, J. (2003). A two-person dynamic equilibrium under ambiguity. *Journal of Economic Dynamics and Control, 27*, 1253–1288.
Epstein, L. G., & Schneider, M. (2007). Learning under ambiguity. *Review of Economic Studies, 74*, 1275–1303.
Epstein, L. G., & Schneider, M. (2008). Ambiguity, information quality and asset pricing. *Journal of Finance, 63*, 197–228.
Epstein, L. G., & Schneider, M. (2010). Ambiguity and asset markets. *Annual Review of Financial Economics, 2*, 315–346.
Epstein, L. G., & Wang, T. (1994). Intertemporal asset pricing under knightian uncertainty. *Econometrica, 62*, 283–322.
Epstein, L. G., & Wang, T. (1995). Uncertainty, risk-neutral measures and security price booms and crashes. *Journal of Economic Theory, 67*, 40–82.
Epstein, L. G., & Zhang, J. (2001). Subjective probabilities on subjectively unambiguous events. *Econometrica, 69*, 265–306.
Ergin, H., & Gul, F. (2009). A theory of subjective compound lotteries. *Journal of Economic Theory, 144*, 899–929.
Ergin, H., & Sarver, T. (2009). *A subjective model of temporal preferences*. Northwestern and WUSTL, Working paper.
Fischhoff, B., & Bruine De Bruin, W. (1999). Fifty–Fifty=50%? *Journal of Behavioral Decision Making, 12*, 149–163.
Fishburn, P. C. (1970). *Utility theory for decision making*. New York: Wiley.
Frisch, R. (1926). Sur un problème d'économie pure. *Norsk Matematisk Forenings Skrifter, 1*, 1–40.
Gajdos, T., Hayashi, T., Tallon, J.-M., & Vergnaud, J.-C. (2008). Attitude toward Imprecise Information. *Journal of Economic Theory, 140*, 27–65.
Gärdenfors, P., & Sahlin, N.-E. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese, 53*, 361–386.
Garlappi, L., Uppal, R., & Wang, T. (2007). Portfolio selection with parameter and model uncertainty: a multi-prior approach. *Review of Financial Studies, 20*, 41–81.
Ghirardato, P. (2002). Revisiting savage in a conditional world. *Economic Theory, 20*, 83–92.
Ghirardato, P., Klibanoff, P., & Marinacci, M. (1998). Additivity with multiple priors. *Journal of Mathematical Economics, 30*, 405–420.
Ghirardato, P., & Marinacci, M. (2001). Risk, ambiguity, and the separation of utility and beliefs. *Mathematics of Operations Research, 26*, 864–890.
Ghirardato, P., & Marinacci, M. (2002). Ambiguity made precise: A comparative foundation. *Journal of Economic Theory, 102*, 251–289.
Ghirardato, P., Maccheroni, F., & Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory, 118*, 133–173.
Ghirardato, P., Maccheroni, F., & Marinacci, M. (2005). Certainty independence and the separation of utility and beliefs. *Journal of Economic Theory, 120*, 129–136.
Ghirardato, P., Maccheroni, F., Marinacci, M., & Siniscalchi, M. (2003). Subjective foundations for objective randomization: A new spin on roulette wheels. *Econometrica, 71*, 1897–1908.
Gilboa, I. (1987). Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics, 16*, 65–88.
Gilboa, I. (2009). *Theory of decision under uncertainty*. Cambridge: Cambridge University Press.

Gilboa, I., Maccheroni, F., Marinacci, M., & Schmeidler, D. (2010). Objective and subjective rationality in a multiple prior model. *Econometrica, 78*, 755–770.
Gilboa, I., Postlewaite, A., & Schmeidler, D. (2008). Probabilities in economic modeling. *Journal of Economic Perspectives, 22*, 173–188.
Gilboa, I., Postlewaite, A., & Schmeidler, D. (2009). Is it always rational to satisfy Savage's axioms? *Economics and Philosophy, 25*(03), 285–296.
Gilboa, I., Postlewaite, A., & Schmeidler, D. (2012). Rationality of belief or: Why Savage's axioms are neither necessary nor sufficient for rationality. *Synthese, 187*(1), 11–31.
Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics, 18*, 141–153. (Working paper, 1986).
Gilboa, I., & Schmeidler, D. (1993). Updating ambiguous beliefs. *Journal of Economic Theory, 59*, 33–49.
Giraud, R. (2005). Objective imprecise probabilistic information, second order beliefs and ambiguity aversion: An axiomatization. In *Proceedings of ISIPTA 2005*, Pittsburgh.
Gollier, C. (2011, forthcoming). Does ambiguity aversion reinforce risk aversion? Applications to portfolio choices and asset pricing. *Review of Economic Studies*.
Greenberg, J. (2000). The right to remain silent. *Theory and Decisions, 48*, 193–204.
Guidolin, M., & Rinaldi, F. (2013). Ambiguity in asset pricing and portfolio choice: A review of the literature. *Theory and Decision, 74*(2), 183–217.
Gul, F., & Pesendorfer, W. (2008). *Measurable ambiguity*. Princeton, Working paper.
Hacking, I. (1975). *The emergence of probability*. Cambridge: Cambridge University Press.
Halevy, Y. (2007). Ellsberg revisited: An experimental study. *Econometrica, 75*, 503–536.
Halevy, Y., & Feltkamp, V. (2005). A bayesian approach to uncertainty aversion. *Review of Economic Studies, 72*, 449–466.
Hanany, E., & Klibanoff, P. (2007). Updating preferences with multiple priors. *Theoretical Economics, 2*, 261–298.
Hanany, E., & Klibanoff, P. (2009). Updating ambiguity averse preferences. *The B.E. Journal of Theoretical Economics, 9*(Advances), Article 37.
Hansen, L. P. (2007). Beliefs, doubts, and learning: Valuing macroeconomic risk. *American Economic Review, 97*, 1–30.
Hansen, L. P., & Sargent, T. J. (2001). Robust control and model uncertainty. *American Economic Review, 91*, 60–66.
Hansen, L. P., & Sargent, T. J. (2008). *Robustness*. Princeton: Princeton University Press.
Hansen, L. P., Sargent, T. J., & Tallarini, T. D. (1999). Robust permanent income and pricing. *Review of Economic Studies, 66*(4), 873–907.
Hansen, L. P., Sargent, T. J., & Wang, N. E. (2002). Robust permanent income and pricing with filtering. *Macroeconomic Dynamics, 6*(01), 40–84.
Hart, S., Modica, S., & Schmeidler, D. (1994). A Neo^2 Bayesian foundation of the maxmin value for two-person zero-SUM games. *International Journal of Game Theory, 23*, 347–358.
Harsanyi, J. C. (1967). Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model. *Management Science, INFORMS, 14*(3), 159–182.
Harsanyi, J. C. (1968). Games with incomplete information played by "Bayesian" players Part II. Bayesian equilibrium points. *Management Science, INFORMS, 14*(5), 320–334.
Hayashi, T., & Miao, J. (2011). Intertemporal substitution and recursive smooth ambiguity preferences. *Theoretical Economics, 6*(3), 423–472.
Hurwicz, L. (1951). Some specification problems and application to econometric models. *Econometrica, 19*, 343–344.
Huygens, C. (1657). *De Ratiociniis in Ludo Aleae*. Amsterdam: van Schooten (trans.: E. D. Sylla, *The art of conjecturing*. Johns Hopkins University Press, 2005).
Jaffray, J.-Y. (1988). Application of linear utility theory to belief functions. In *Uncertainty and intelligent systems* (pp. 1–8). Berlin: Springer.
Jaffray, J. Y. (1989). Coherent bets under partially resolving uncertainty and belief functions. *Theory and Decision, 26*(2), 99–105.
Ju, N., & Miao, J. (2012). Ambiguity, learning, and asset returns. *Econometrica, 80*(2), 559–591.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.
Kajii, A., & Ui, T. (2006). Agreeable bets with multiple priors. *Journal of Economic Theory, 128*, 299–305.
Kajii, A., & Ui, T. (2009). Interim efficient allocations under uncertainty. *Journal of Economic Theory, 144*, 337–353.
Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association, 91*, 1343–1370.
Keynes, J. M. (1921). *A treatise on probability*. London: MacMillan.
Keynes, J. M. (1937). *The Quarterly Journal of Economics*. From *The Collected Writings of John Maynard Keynes* (Vol. XIV, pp. 109–123).
Klibanoff, P. (2001a). Stochastically independent randomization and uncertainty aversion. *Economic Theory, 18*, 605–620.
Klibanoff, P. (2001b). Characterizing uncertainty aversion through preference for mixtures. *Social Choice and Welfare, 18*, 289–301.
Klibanoff, P., Marinacci, M., & Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica, 73*, 1849–1892.
Klibanoff, P., Marinacci, M., & Mukerji, S. (2009). Recursive smooth ambiguity preferences. *Journal of Economic Theory, 144*, 930–976.
Knight, F. H. (1921). *Risk, uncertainty, and profit*. Boston/New York: Houghton Mifflin.
Kobberling, V., & Wakker, P. P. (2003). Preference foundations for nonexpected utility: A generalized and simplified technique. *Mathematics of Operations Research, 28*, 395–423.
Kocherlakota, N. R. (2007). Model fit and model selection. *Federal Reserve Bank of St. Louis Review, 89*, 349–360.
Kopylov, I. (2001). *Procedural rationality in the multiple prior model*. Rochester, Working paper.
Kopylov, I. (2010). *Simple axioms for countably additive subjective probability*. UC Irvine, Working paper.
Kreps, D. M. (1979). A representation theorem for "preference for flexibility". *Econometrica: Journal of the Econometric Society, 47*(3), 565–577.
Kreps, D. (1988). *Notes on the theory of choice* (Underground classics in economics). Boulder: Westview Press.
Laplace, P. S. (1814). *Essai Philophique sur les Probabilites*. Paris: Gauthier-Villars (English ed., 1951, A philosophical essay on probabilities. New York: Dover).
Levi, I. (1974). On indeterminate probabilities. *Journal of Philosophy, 71*, 391–418.
Levi, I. (1980). *The enterprise of knowledge*. Cambridge: MIT.
Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability*. Berkeley/Los Angeles: University of California Press.
Lipman, B., & Pesendorfer, W. (2013). Temptation. In D. Acemoglu, M. Arellano, & E. Dekel (Eds.), *Advances in economics and econometrics: Theory and applications.* Cambridge: Cambridge University Press.
Maccheroni, F., & Marinacci, M. (2005). A strong law of large numbers for capacities. *The Annals of Probability, 33*, 1171–1178.
Maccheroni, F., Marinacci, M., & Rustichini, A. (2006a). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica, 74*, 1447–1498.
Maccheroni, F., Marinacci, M., & Rustichini, A. (2006b). Dynamic variational preference. *Journal of Economic Theory, 128*, 4–44.
Machina, M. J. (1982). 'Expected Utility' analysis without the independence axiom. *Econometrica, 50*, 277–323.
Machina, M. J. (2004). Almost-objective uncertainty. *Economic Theory, 24*, 1–54.
Machina, M. J. (2005). 'Expected Utility/Subjective Probability' analysis without the sure-thing principle or probabilistic sophistication. *Economic Theory, 26*, 1–62.
Machina, M. J., & Schmeidler, D. (1992). A more robust definition of subjective probability. *Econometrica, 60*, 745–780.

Marinacci, M. (1999). Limit laws for non-additive probabilities and their frequentist interpretation. *Journal of Economic Theory, 84*, 145–195.
Marinacci, M. (2002a). Probabilistic sophistication and multiple priors. *Econometrica, 70*, 755–764.
Marinacci, M. (2002b). Learning from ambiguous urns. *Statistical Papers, 43*, 143–151.
Marinacci, M., & Montrucchio, L. (2004). Introduction to the mathematics of ambiguity. In I. Gilboa (Ed.), *Uncertainty in economic theory*. New York: Routledge.
Miao, J. (2004). A note on consumption and savings under knightian uncertainty. *Annals of Economics and Finance, 5*, 299–311.
Miao, J. (2009). Ambiguity, risk and portfolio choice under incomplete information. *Annals of Economics and Finance, 10*, 257–279.
Miao, J., & Wang, N. (2011). Risk, uncertainty, and option exercise. *Journal of Economic Dynamics and Control, 35*(4), 442–461.
Milnor, J. (1954). Games against nature. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision processes*. New York: Wiley.
Montesano, A., & Giovannone, F. (1996). Uncertainty aversion and aversion to increasing uncertainty. *Theory and Decision, 41*, 133–148.
Mukerji, S. (1998). Ambiguity aversion and the incompleteness of contractual form. *American Economic Review, 88*, 1207–1232.
Mukerji, S. (2009). Foundations of ambiguity and economic modelling. *Economics and Philosophy, 25*, 297–302.
Mukerji, S., & Tallon, J.-M. (2001). Ambiguity aversion and incompleteness of financial markets. *Review of Economic Studies, 68*, 883–904.
Mukerji, S., & Tallon, J.-M. (2004). An overview of economic applications of David Schmeidler's models of decision making under uncertainty. In I. Gilboa (Ed.), *Uncertainty in economic theory*. New York: Routledge.
Nakamura, Y. (1990). Subjective expected utility with non-additive probabilities on finite state spaces. *Journal of Economic Theory, 51*, 346–366.
Nau, R. F. (2001, 2006). Uncertainty aversion with second-order utilities and probabilities. *Management Science, 52*, 136–145. (see also *Proceedings of ISIPTA 2001*).
Nau, R. (2011). Risk, ambiguity, and state-preference theory. *Economic Theory, 48*(2–3), 437–467.
Nehring, K. (1999). Capacities and probabilistic beliefs: A precarious coexistence. *Mathematical Social Sciences, 38*, 197–213.
Nehring, K. (2001). Common priors under incomplete information: A unification. *Economic Theory, 18*(3), 535–553.
Nishimura, K., & Ozaki, H. (2004). Search and knightian uncertainty. *Journal of Economic Theory, 119*, 299–333.
Nishimura, K., & Ozaki, H. (2007). Irreversible investment and knightian uncertainty. *Journal of Economic Theory, 136*, 668–694.
Olszewski, W. B. (2007). Preferences over sets of lotteries. *Review of Economic Studies, 74*, 567–595.
Ore, O. (1960). Pascal and the invention of probability theory. *American Mathematical Monthly, 67*, 409–419.
Ortoleva, P. (2010). Status quo bias, multiple priors and uncertainty aversion. *Games and Economic Behavior, 69*, 411–424.
Ozdenoren, E., & Peck, J. (2008). Ambiguity aversion, games against nature, and dynamic consistency. *Games and Economic Behavior, 62*, 106–115.
Pascal, B. (1670). *Pensées sur la Religion et sur Quelques Autres Sujets*.
Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence, 29*, 241–288.
Pires, C. P. (2002). A rule for updating ambiguous beliefs. *Theory and Decision, 33*, 137–152.
Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behaviorand Organization, 3*, 225–243.

Ramsey, F. P. (1926a). Truth and probability. In R. Braithwaite (Ed.), *The foundation of mathematics and other logical essays*, (1931). London: Routledge and Kegan.
Ramsey, F. P. (1926b). Mathematical logic. *Mathematical Gazette, 13*, 185–194.
Rigotti, L., & Shannon, C. (2005). Uncertainty and risk in financial markets. *Econometrica, 73*, 203–243.
Rigotti, L., Shannon, C., Strzalecki, T. (2008). Subjective beliefs and ex ante trade. *Econometrica, 76*, 1167–1190.
Rosenmueller, J. (1971). On core and value. *Methods of Operations Research, 9*, 84–104.
Rosenmueller, J. (1972). Some properties of convex set functions, Part II. *Methods of Operations Research, 17*, 287–307.
Saito, K. (2015). Preferences for flexibility and randomization under uncertainty. *The American Economic Review, 105*(3), 1246–1271.
Sarin, R., & Wakker, P. P. (1992). A simple axiomatization of nonadditive expected utility. *Econometrica, 60*, 1255–1272.
Sarin, R., & Wakker, P. P. (1998). Dynamic choice and nonexpected utility. *Journal of Risk and Uncertainty, 17*, 87–119.
Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley. (2nd ed. in 1972, Dover)
Schmeidler, D. (1986). Integral representation without additivity. *Proceedings of the American Mathematical Society, 97*, 255–261.
Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica, 57*, 571–587. (Working paper, 1982).
Seo, K. (2009). Ambiguity and second-order belief. *Econometrica, 77*, 1575–1605.
Segal, U. (1987). The ellsberg paradox and risk aversion: An anticipated utility approach. *International Economic Review, 28*, 175–202.
Segal, U. (1990). Two-stage lotteries without the reduction axiom. *Econometrica, 58*, 349–377.
Seidenfeld, T., & Wasserman, L. (1993). Dilation for sets of probabilities. *The Annals of Statistics, 21*, 1139–1154.
Shafer, G. (1976). *A mathematical theory of evidence*. Princeton: Princeton University Press.
Shafer, G. (1986). Savage revisited. *Statistical Science, 1*, 463–486.
Shapley, L. S. (1972). Cores of convex games. *International Journal of Game Theory, 1*, 11–26. (Working paper, 1965).
Siniscalchi, M. (2006a). A behavioral characterization of plausible priors. *Journal of Economic Theory, 128*, 91–135.
Siniscalchi, M. (2006b). Dynamic choice under ambiguity. *Theoretical Economics, 6*(3). September 2011.
Siniscalchi, M. (2009a). Vector expected utility and attitudes toward variation. *Econometrica, 77*, 801–855.
Siniscalchi, M. (2009b). Two out of three ain't bad: A comment on 'The ambiguity aversion literature: A critical assessment'. *Economics and Philosophy, 25*, 335–356.
Smith, C. A. B. (1961). Consistency in statistical inference and decision. *Journal of the Royal Statistical Society, Series B, 23*, 1–25.
Stinchcombe, M. (2003). *Choice and games with ambiguity as sets of probabilities*. UT Austin, Working paper.
Strzalecki, T. (2010, forthcoming). Axiomatic foundations of multiplier preferences. *Econometrica*.
Suppe, F. (1977). *The structure of scientific theories*. Champaign: University of Illinois Press.
Treich, N. (2010). The value of a statistical life under ambiguity aversion. *Journal of Environmental Economics and Management, 59*, 15–26.
Tversky, A., & Fox, C. (1995). Weighing risk and uncertainty. *Psychological Review, 102*, 269–283.
Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*, 297–323.
van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford: Oxford University Press.

Viero, M.-L. (2009) Exactly what happens after the Anscombe-Aumann race? Representing preferences in vague environments. *Economic Theory, 41*, 175–212.
von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton: Princeton University Press.
Wakker, P. P. (1989a). Continuous subjective expected utility with nonadditive probabilities. *Journal of Mathematical Economics, 18*, 1–27.
Wakker, P. P. (1989b). *Additive representations of preferences: A new foundation of decision analysis*. Dordrecht: Kluwer.
Wakker, P. P. (1990). Characterizing optimism and pessimism directly through comonotonicity. *Journal of Economic Theory, 52*, 453–463.
Wakker, P. P. (1991). Testing and characterizing properties of nonadditive measures through violations of the sure-thing principle. *Econometrica, 69*, 1039–1059.
Wakker, P. P. (2010). *Prospect theory*. Cambridge: Cambridge University Press.
Wald, A. (1950). *Statistical decision functions*. New York: Wiley.
Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. London: Chapman and Hall.
Wang, T. (2003a). *A class of multi-prior preferences*. UBC, Working paper.
Wang, T. (2003b). Conditional preferences and updating. *Journal of Economic Theory, 108*, 286–321.
Welch, B. L. (1939). On confidence limits and sufficiency, and particular reference to parameters of location. *Annals of Mathematical Statistics, 10*, 58–69.
Yaari, M. E. (1969). Some remarks on measures of risk aversion and on their uses. *Journal of Economic Theory, 1*, 315–329.
Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica, 55*, 95–115.
Zhang, J. (2002). Subjective ambiguity, expected utility, and choquet expected utility. *Economic Theory, 20*, 159–181.

Chapter 22
State-Dependent Utilities

Mark J. Schervish, Teddy Seidenfeld, and Joseph B. Kadane

Introduction

Expected utility theory is founded upon at least one of several axiomatic derivations of probabilities and utilities from expressed preferences over acts, Savage (1954), deFinetti (1974), Anscombe and Aumann (1963), and Ramsey (1926). These theories provide for the simultaneous existence of a unique personal probability over the states of nature and a unique (up to positive affine transformations) utility function over the prizes such that the ranking of acts is by expected utility. For example, suppose that there are n states of nature which form the set $S = \{s_1, \ldots, s_n\}$ and m prizes in the set $Z = \{z_1, \ldots, z_m\}$. An example of an act is a function f mapping S to Z. That is, if $f(s_i) = z_j$, then we receive prize z_j if state s_i occurs. (We will consider more complicated acts than this later.) Now, suppose that

This research was reported, in part, at the Indo-United States Workshop on Bayesian Analysis in Statistics and Econometrics. The research was supported by National Science Foundation grants DMS-8805676 and DMS-8705646, and Office of Naval Research contract N00014-88-K0013. The authors would like to thank Morris DeGroot, Bruce Hill, Irving LaValle, Isaac Levi, and Herman Rubin for helpful comments during the preparation of this paper. We especially thank the associate editor for the patience and care that was given to this submission.

M.J. Schervish (✉)
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: mark@cmu.edu

T. Seidenfeld
Departments of Philosophy and Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA

J.B. Kadane
Departments of Statistics and Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_22

there is a probability over the states such that $p_i = \Pr(s_i)$ and that there is a utility U over prizes. To say that acts are ranked by expected utility means that we strictly prefer act g to act f if and only if

$$\sum_{i=1}^{n} p_i U(f(s_i)) < \sum_{i=1}^{n} p_i U(g(s_i)). \tag{22.1}$$

If we allow the utilities of prizes to vary conditionally on which state of nature occurs, we can rewrite (22.1) as

$$\sum_{i=1}^{n} p_i U_i(f(s_i)) < \sum_{i=1}^{n} p_i U_i(g(s_i)), \tag{22.2}$$

where $U_i(z_j)$ is the utility of prize z_j given that state s_i occurs. However, without restrictions on the degree to which U_i can differ from $U_{i'}$ for $i \neq i'$, the uniqueness of the personal probability no longer holds. For example, let $q_1, \ldots, q_n$ be another probability over the states such that $p_i > 0$ if and only if $q_i > 0$. Then, for an arbitrary act f, $\sum_{i=1}^{n} q_i V_i(f(s_i)) = \sum_{i=1}^{n} p_i U_i(f(s_i))$, where $V_i(\cdot) = p_i U_i(\cdot)/q_i$ when $q_i > 0$ (V_i can be arbitrary when $q_i = 0$.) In this case, it is impossible to determine an agent's personal probability by studying the agent's preferences for acts. Rubin (1987) notes this fact and develops an axiom system that does not lead to a separation of probability and utility. Arrow (1974) considers the problem for insurance. A footnote in Arrow (1974) credits Herman Rubin with raising this same issue in an unpublished 1964 lecture.

DeGroot (1970) begins his derivation of expected utility theory by assuming that the concept of "at least as likely as" is an undefined primitive. This allows a construction of probability without reference to preferences. However, DeGroot also needs to introduce preferences among acts in order to derive a utility function. In section "State-Independent Utility", we will examine the axiomatization of VonNeumann and Morgenstern (1947) together with the extension of Anscombe and Aumann (1963) to see how it attempts to avoid the non-uniqueness problem just described. In section "Savage's Postulates", we look at the system of Savage (1954) with the same goal in mind. In section "An Example", we give an example that illustrates that the problem can still arise despite the best efforts of those who have derived the theories. This example leads to a critical examination of the theory of deFinetti (1974) in section "deFinetti's Gambling Approach". While reviewing an example from Savage (1954) in section "Savage's 'Small Worlds' Example", we see how close Savage was to discovering the non-uniqueness problem in connection with his own theory. In section "How to Elicit Unique Probabilities and Utilities Simultaneously", we describe a method for obtaining a unique personal probability and state-dependent utility based on a proposal of Karni et al. (1983).

State-Independent Utility

Following VonNeumann and Morgenstern (1947), we generalize the concept of act introduced in section "Introduction" by allowing randomization. That is, suppose that the agent is comfortable declaring probabilities for an auxiliary experiment the results of which he/she believes would, in no way, alter his/her preferences among acts. Furthermore, assume that this auxiliary experiment has events with arbitrary probabilities (for example, it may produce a random variable with continuous distribution). Define a *lottery* as follows. If $A_1, \ldots, A_m$ is a partition of the possible outcomes of the auxiliary experiment with $\alpha_j = \Pr(A_j)$ for each j, then the lottery $(\alpha_1, \ldots, \alpha_m)$ awards prize z_j if A_j occurs. We assume that the choice of the partition events $A_1, \ldots, A_m$ does not affect the lottery. That is, if $B_1, \ldots, B_m$ is another partition such that $\Pr(B_j) = \alpha_j$ for each j also, then the lottery that awards prize z_j when B_j occurs is, to the agent, *the same lottery* as the one described above in terms of the A_j. In this way, a lottery is just a simple probability distribution over prizes that is independent of the state of nature. Any two lotteries that award the prizes with the same probabilities are considered the same lottery. If L_1 and L_2 are two lotteries $(\alpha_1, \ldots, \alpha_m)$ and $(\beta_1, \ldots, \beta_m)$ respectively, then for $0 \leq \lambda \leq 1$, we denote by $\lambda L_1 + (1-\lambda)L_2$ the lottery $(\lambda\alpha_1 + (1-\lambda)\beta_1, \ldots, \lambda\alpha_m + (1-\lambda)\beta_m)$.

If we consider only lotteries, we can introduce some axioms for preferences among lotteries. For convenience, we will henceforth assume that there exist two lotteries such that the agent has a strict preference for one over the other. Otherwise, the preference relation is trivial and no interesting results are obtained.

Axiom 1 (Weak Order). There is a weak order $\preceq$ among lotteries such that $L_1 \preceq L_2$ if and only if L_1 is not strictly preferred to L_2.

This axiom requires that weak preference among lotteries be transitive, reflexive, and connected. If we define *equivalence* to mean "no strict preference in either direction", then equivalence is transitive also.

Definition 1. *Assuming Axiom 1, we say that* L_1 is equivalent to L_2 *(denoted* $L_1 \sim L_2$*) if* $L_1 \preceq L_2$ *and* $L_2 \preceq L_1$. *We say* L_2 *is* strictly preferred to L_1 *(denoted* $L_1 \prec L_2$*) if* $L_1 \preceq L_2$ *but not* $L_2 \preceq L_1$.

The axiom that does most of the work is one that entails stochastic dominance.

Axiom 2 (Independence). For each L, L_1, L_2, and $0 < \alpha < 1$, $L_1 \preceq L_2$ if and only if $\alpha L_1 + (1-\alpha)L \preceq \alpha L_2 + (1-\alpha)L$.

A third axiom is often introduced to guarantee that utilities are real valued.

Axiom 3 (Archimedean). If $L_1 \prec L_2 \prec L_3$, then there exists $0 < \alpha < 1$ such that $L_2 \sim (1-\alpha)L_1 + \alpha L_3$.

Axiom 3 prevents L_3 from being infinitely better than L_2 and it prevents L_1 from being infinitely worse than L_2.

With axioms equivalent to the three above, VonNeumann and Morgenstern (1947) prove that there exists a utility over prizes U such that $(\alpha_1, \ldots, \alpha_m) \preceq (\beta_1, \ldots, \beta_m)$ if and only if $\sum_{i=1}^m \alpha_i U(z_i) \leq \sum_{i=1}^m \beta_i U(z_i)$. This utility is unique up to positive affine transformation. In fact, it is quite easy (and useful for the example in section "An Example") to construct the utility function from the stated preferences. Pick an arbitrary pair of lotteries L_0 and L_1 such that $L_0 \prec L_1$. Assign these lotteries $U(L_0) = 0$ and $U(L_1) = 1$. For all other lotteries L, the utilities are assigned as follows. If $L_0 \preceq L \preceq L_1$, $U(L)$ is that α such that $(1-\alpha)L_0 + \alpha L_1 \sim L$. If $L \prec L_0$, then $U(L) = -\alpha/(1-\alpha)$, where $(1-\alpha)L + \alpha L_1 \sim L_0$ (hence $\alpha \neq 1$.) If $L_1 \prec L$, then $U(L) = 1/\alpha$, where $(1-\alpha)L_0 + \alpha L \sim L_1$. The existence of these α values is guaranteed by Axiom 3 and their uniqueness follows from Axiom 2.

To be able to handle acts in which the prizes vary with the state of nature, Anscombe and Aumann (1963) introduced a fourth axiom, which was designed to say that the preferences among prizes did not vary with the state. Before stating this axiom, we introduce a more general act, known as a *horse lottery*.

Definition 2. *A function mapping states of nature to lotteries is called a* horse lottery.

That is, if H is a horse lottery such that $H(s_i) = L_i$ for each i, then, if state s_i occurs, the prize awarded is the prize that lottery L_i awards. The L_i can be all different or some (or all) of them can be the same. If $H(s_i) = L$ for all i, then we say $H = L$. If H_1 and H_2 are two horse lotteries such that $H_j(s_i) = L_i^{(j)}$ for each i and j, and $0 \leq \alpha \leq 1$, then we denote by $\alpha H_1 + (1-\alpha)H_2$ the horse lottery H such that $H(s_i) = \alpha L_1^{(i)} + (1-\alpha)L_2^{(i)}$ for each i. Axioms 1 and 2, when applied to preferences among horse lotteries, imply that the choice of an act has no effect on the probabilities of the state of nature.

Definition 3. *A state of nature s_i is called* null *if, for each pair of horse lotteries H_1 and H_2 satisfying $H_1(s_j) = H_2(s_j)$ for all $j \neq i$, $H_1 \sim H_2$. A state is called* non-null *if it is not null.*

Axiom 4 (State-independence). For each non-null state s_i, each pair of lotteries (L_1, L_2), and each pair of horse lotteries H_1 and H_2 satisfying $H_1(s_j) = H_2(s_j)$ for $j \neq i$, $H_1(s_i) = L_1$, and $H_2(s_i) = L_2$, we have $L_1 \prec L_2$ if and only if $H_1 \prec H_2$.

Axiom 4 says that a strict preference between two lotteries is reproduced for every pair of horse lotteries that differ only in some non-null state, and their difference in that state is that each of them equals one of the two lotteries. With this setup, Anscombe and Aumann (1963) prove the following theorem.

Theorem 1 (Anscombe and Aumann). *Under Axioms 1, 2, 3, and 4, there exist a unique probability P over the states and utility U over prizes (unique up to positive affine transformation) such that $H_1 \preceq H_2$ if and only if $\sum_{i=1}^n P(s_i)U(H_1(s_i)) \leq \sum_{i=1}^n P(s_i)U(H_2(s_i))$, where, for each lottery $L = (\alpha_1, \ldots, \alpha_m)$, $U(L)$ stands for $\sum_{j=1}^m \alpha_j U(z_j)$.*

Even when the four axioms above hold, there is no requirement that the utility function U be the same conditional on each state of nature. As we did when we constructed Eq. (22.2), we could allow $U_i(z_j) = a_i U(z_j) + b_i$ where each $a_i > 0$. Then we could let $Q(s_i) = a_i P(s_i) / \sum_{k=1}^{n} a_k P(s_k)$. It would now be true that $H_1 \preceq H_2$ if and only if $\sum_{i=1}^{n} Q(s_i) U_i(H_1(s_i)) \leq \sum_{i=1}^{n} Q(s_i) U_i(H_2(s_i))$. The uniqueness of the probability in Theorem 1 depends on the use of a state-independent utility U. Hence, one cannot determine an agent's probability from the agent's stated preferences unless one assumes that the agent's utility is state-independent. This may not seem like a serious difficulty when Axiom 4 holds. However, as we will see in section "An Example", the problem is more complicated.

Savage's Postulates

Savage (1954) gives a set of postulates that does not rely on an auxiliary randomization in order to extract probabilities and utilities from preferences. Rather, the postulates rely on the use of prizes that can be considered as "constant" across states. Savage's most general acts are functions from states to prizes. Because he does not introduce an auxiliary randomization, he requires that there be infinitely many states. The important features of Savage's theory, for this discussion, are the first three postulates and a few definitions. Some of the axioms and definitions are stated in terms of *events*, which are sets of states. The postulates of Savage are consistent with the axioms of section "State-Independent Utility" in that they both provide models for preference by maximizing expected utility.

The first postulate of Savage is the same as Axiom 1. The second postulate requires a definition of conditional preference.

Definition 4. *Let B be an event. We say that $f \preceq g$* given *B if and only if*

- *$f' \preceq g'$ for each pair f' and g' such that $f'(s) = f(s)$ for all $s \in B$, $g'(s) = g(s)$ for all $s \in B$, and $f'(s) = g'(s)$ for all $s \notin B$.*
- *and $f' \preceq g'$ for every such pair or for none.*

The second postulate is an analog of Axiom 2. (See Fisburn 1970, p. 193.)

Postulate 2. *For each pair of acts f and g and each event B, either $f \preceq g$ given B or $g \preceq f$ given B.*

Savage has a concept of *null event* that is similar to the concept of null state from Definition 3.

Definition 5. *An event B is* null *if, for every pair of acts f and g, $f \preceq g$ given B. An event B is* non-null *if it is not null.*

The third postulate of Savage concerns acts that are constant such as $f(s) = z$ for all s, where z is a single prize. For convenience, we will call such an act f by the name z also.

Postulate 3. *For each non-null event B and each pair of prizes z_1 and z_2 (considered as constant acts), $z_1 \preceq z_2$ if and only if $z_1 \preceq z_2$ given B.*

Savage's definition of probability relies on Postulate 3.

Definition 6. *Suppose that A and B are events. We say that A* is at least as likely as *B if, for each pair of prizes z and w with $z \prec w$ we have $f_B \preceq f_A$, where $f_A(s) = w$ if $s \in A$, $f_A(s) = z$ if $s \notin A$, $f_B(s) = w$ if $s \in B$, and $f_B(s) = z$ if $s \notin B$.*

Postulate 2 guarantees that, with f_A and f_B as defined in Definition 6, either $f_B \preceq f_A$ no matter which pair of prizes z and w one chooses (so long as $z \prec w$) or $f_A \preceq f_B$ no matter which pair of prizes one chooses.

Postulate 3 says that the *relative* values of prizes cannot change between states. Savage (1954, p. 25) suggests that problems in locating prizes which satisfy this postulate might be solved by a clever redescription. For example, rather than describing prizes as "receiving a bathing suit" and "receiving a tennis racket" (whose relative values change depending on which of the two states "picnic at the beach" or "picnic in the park" occurs), Savage suggests that the prizes might be "a refreshing swim with friends," "sitting alone on the beach with a tennis racket," etc. However, we do not see how to carry out such redescriptions while satisfying Savage's structural assumption that each prize is available as an outcome under each state. (What does it mean to receive the prize "sitting alone on the beach with a tennis racket" when the state "picnic in the park" occurs?)

Our problem, however, is deeper than this. Definition 6 assumes that the *absolute* values of prizes do not change from state to state. For example, suppose that A and B are disjoint and the value of z is 1 for the states in A and 2 for the states in B. Similarly, suppose that the value of w is 2 for the states in A and 4 for the states in B. Then, even if A is more likely than B, but is not twice as likely, we would get $f_A \prec f_B$, and we would conclude, by Definition 6, that B is more likely than A. The example in section "An Example" (using just one of the currencies), as well as our interpretation of Savage's "small worlds" problem (in section "Savage's 'Small Worlds' Example") suggest that it might be very difficult to find prizes with the property that their "absolute" values do not change from state to state even though their "relative" values remain the same from state to state.

deFinetti's Gambling Approach

deFinetti (1974) assumes that there is a set of prizes with numerical values such that utility is linear in the numerical value. That is, a prize numbered 4 is worth twice as much as a prize numbered 2. More specifically, to say that utility is linear in the numerical values of prizes, we mean the following. For each pair of prizes, (z_1, z_2) with $z_1 < z_2$, and each $0 \le \alpha \le 1$, the lottery that pays z_1 with probability $1 - \alpha$ and pays z_2 with probability α (using the auxiliary randomization of section "State-Independent Utility") is equivalent to the lottery that pays prize

$(1-\alpha)z_1 + \alpha z_2$ for sure. Using such a set of prizes, deFinetti supposes that an agent will accept certain gambles that pay these prizes. If f is an act, to gamble on f means to accept a contract that pays the agent the prize $c(f(s) - x)$ when state s occurs, where c and x are some values. A negative outcome means that the agent has to pay out, while a positive outcome means that the agent gains some amount.

Definition 7. *The* prevision *of an act f is the number x that one would choose so that all gambles of the form $c(f - x)$ would be accepted, for all small values of c, both positive and negative.*

If an agent is willing to gamble on each of several acts, then it is assumed that the agent will also gamble on them simultaneously. (For a critical discussion of this point, see Kadane and Winkler (1988) and Schick (1986).)

Definition 8. *A collection of previsions for acts is* coherent *if, for each finite set of the acts, say $f_1, \ldots, f_n$ with previsions $x_1, \ldots, x_n$ respectively, and each set of numbers $c_1, \ldots, c_n$, we have* $\sup_{all\ s} \sum_{i=1}^{n} c_i(f_i(s) - x_i) \geq 0$. *Otherwise, the previsions are* incoherent.

deFinetti (1974) proves that a collection of previsions of bounded acts is coherent if and only if there exists a finitely additive probability such that the prevision of each act is its expected value. This provides a method of eliciting probabilities by asking an agent to specify previsions for acts such as $f(s) = 1$ if $s \in A$ and $f(s) = 0$ if $s \notin A$. The prevision of such an act f would be its probability if the previsions are coherent. As plausible as this sounds, the example below casts doubt on the ability of deFinetti's program to elicit probabilities accurately.

An Example

Let the set of available prizes be various amounts of Dollars. We suppose that there are three states of nature, which we will describe in more detail later, and we suppose that the agent expresses preferences that satisfy the axioms of section "State-Independent Utility" and the postulates of Savage (1954). Furthermore, suppose that the agent's utility for money is linear. That is, for each state i, $U_i(\$cx) = cU_i(\$x)$. In particular, $U_i(\$0) = 0$. Now, we offer the agent three horse lotteries H_1, H_2, and H_3 whose outcomes are

	State of Nature		
	s_1	s_2	s_3
H_1	\$1	\$0	\$0
H_2	\$0	\$1	\$0
H_3	\$0	\$0	\$1

Suppose that the agent claims that these three horse lotteries are equivalent. If we assume that the agent has a state-independent utility, the expected utility of H_i is $U(\$1)P(s_i)$. It follows from the fact that the three horse lotteries are equivalent, that $P(s_i) = 1/3$ for each i.

Next, we alter the set of prizes to be various Yen amounts (the Japanese currency). Suppose that we offer the agent three Yen horse lotteries H_4, H_5, and H_6 whose outcomes are

	State of Nature		
	s_1	s_2	s_3
H_4	100Y	0Y	0Y
H_5	0Y	125Y	0Y
H_6	0Y	0Y	150Y

If the agent were to claim that these three horse lotteries were equivalent, and if we assumed that the agent used a state-independent utility for Yen prizes, then $P(s_1)U(100\text{Y}) = P(s_2)U(125\text{Y}) = P(s_3)U(150\text{Y})$. Supposing that the agent's utility is linear in Yen, as it was in dollars, we conclude that $P(s_1) = 1.25P(s_2) = 1.5P(s_3)$. It follows that $P(s_1) = .4054$, $P(s_2) = .3243$, and $P(s_3) = .2703$. It would seem incoherent for the agent to express both sets of equivalences since it appears that the agent is now committed to two different probability distributions over the three states. This is not correct, however, as we now see.

Suppose that the three states of nature represent three different exchange rates between Dollars and Yen. $s_1 = \{\$1 \text{ is worth } 100\text{Y}\}$, $s_2 = \{\$1 \text{ is worth } 125\text{Y}\}$, and $s_3 = \{\$1 \text{ is worth } 150\text{Y}\}$. Suppose further that the agent can change monetary units at the prevailing rate of exchange without any penalty. As far as this agent is concerned, H_i and H_{3+i} are worth exactly the same for $i = 1, 2, 3$ since, in each state the prizes they award are worth the same amount. The problem that arises in this example is that the two probability distributions were constructed under incompatible assumptions. The discrete uniform probability was constructed under the assumption that $U(\$1)$ is the same in all three states, while the other probability was constructed under the assumption that $U(100\text{Y})$ was the same in all three states. Clearly these cannot both be true given the nature of the states. What saves both Theorem 1 and Savage's theory is that preference can be represented by expected utility *no matter which of the two assumptions one makes*. Unfortunately, this same fact makes the uniqueness of the probability relative to the choice of which prizes count as constants in terms of utility. There are two different representations of the agent's preferences by probability and state-independent utility. But what is state-independent in one representation is state-dependent in the other.

If we allow both types of prizes at once, we can calculate the marginal exchange rate for the agent. That is, we can ask, "For what value x will the agent claim that \$1 and xY are equivalent?" This question can be answered using either of the two probability-utility representations and the answers will be the same. First, with

Dollars having constant value, the expected utility of a horse lottery paying \$1 in all three states is $U(\$1)$. The expected value of the horse lottery paying xY in all three states is

$$\begin{aligned}\frac{U_1(x\text{Y}) + U_2(x\text{Y}) + U_3(x\text{Y})}{3} &= \frac{1}{3}\left(\frac{x}{100}U(\$1) + \frac{x}{125}U(\$1) + \frac{x}{150}U(\$1)\right)\\ &= .008222xU(\$1),\end{aligned}$$

using the linearity of utility and the state-specific exchange rates. By setting this expression equal to $U(\$1)$, we obtain that $x = 121.62$. Equivalently, we can calculate the exchange rate assuming that Yen have constant value over states. The act paying xY in all states has expected utility $U(x\text{Y}) = .01xU(100\text{Y})$. The act paying \$1 in all states has expected utility

$$\begin{aligned}&.4054U_1(\$1) + .3243U_2(\$1) + .2703U_3(\$1)\\ &= .4054U(100\text{Y}) + .3243U(125\text{Y}) + .2703U(150\text{Y})\\ &= U(100\text{Y})[.4054 + .3243 \times 1.25 + .2703 \times 1.5]\\ &= 1.2162U(100\text{Y}).\end{aligned}$$

Setting this equal to $.01xU(100\text{Y})$ yields $x = 121.62$, which is the same exchange rate as calculated earlier.

The implications of this example for elicitation are staggering. Suppose we attempt to elicit the agent's probabilities over the three states by offering acts in Dollar amounts using deFinetti's gambling approach from section "deFinetti's Gambling Approach". The agent has utility that is linear in both Dollars and Yen without reference to the states, hence deFinetti's program will apply. To see this, select two prizes, such as \$0 and \$1, to have utilities 0 and 1 respectively. Then, for $0 < x < 1$, $U(\$x)$ must be the value c that makes the following two lotteries equivalent: $L_1 = \$x$, for certain and $L_2 = \$1$ with probability c and \$0 with probability $1 - c$. Assuming that Dollars have constant utility, it is obvious that $c = x$. Assuming that Yen have constant utility, the expected utility of L_1 is $1.2162xU(100\text{Y})$ and the expected utility of L_2 is $cU(121.62Y)$. These two are the same if and only if $x = c$. A similar argument works for x not between 0 and 1, and a similar argument works when the two prizes with utilities 0 and 1 are Yen prizes. Now, suppose that the agent actually uses the state-independent utility for Dollars and the discrete uniform distribution to rank acts, but the eliciter does not know this. The eliciter will try to elicit the agent's probabilities for the states by offering gambles in Yen (linear in utility). For example, the agent claims that the gamble $c(f - 40.54)$ would be accepted for all small values of c, where $f(s) = 150\text{Y}$ if $s = s_3$ and equals 0Y otherwise. The reason for this is that, since 150Y equals \$1 when s_3 occurs, the winnings are \$1 when s_3 occurs, which has probability 1/3. The marginal exchange rate is 121.62Y for \$1, so the appropriate amount to pay (no matter which state occurs), in order to win \$1 when s_3 occurs, is \$1/3, which

equals $121.62\text{Y}/3 = 40.54\text{Y}$. Realizing that utility is linear in Yen, the eliciter now decides that $\Pr(s_3)$ must equal $40.54/150 = .2703$. Hence, the eliciter elicits the wrong probability, even though the agent is coherent!

The expressed preferences satisfy the four axioms of section "State-Independent Utility", all of Savage's postulates, and deFinetti's linearity condition, but we are still unable to determine the probabilities of the states based only on preferences. The problem becomes clearer if we allow both Dollar and Yen prizes at the same time. Now, it is impossible for a single utility to be state-independent for all prizes. That is, Axiom 4 and Postulate 3 would no longer hold. Things are more confusing in deFinetti's framework, because there is no room for state-dependent utilities. The agent would appear to have two different probabilities for the same event even though there would be no incoherency.

Savage's 'Small Worlds' Example

In Section 5.5 of Savage (1954), the topic of *small worlds* is discussed. An anomaly occurs in this discussion, and Savage seems to imply that it is an effect of the construction of the small world. In this section, we briefly introduce small worlds and then explain why we believe that the anomaly discovered by Savage is actually another example of the non-uniqueness illustrated in section "An Example". The fact that it arose in the discussion of small worlds is a mere coincidence. We show how precisely the same effect arises without any mention of small worlds.

A small world can be thought of as a description of the states of nature in which each state can actually be partitioned into several smaller states, but we don't actually do the partitioning when making comparisons between acts. For a mathematical example, Savage mentions the following case. Consider the unit square $S = \{(x, y) : 0 \le x, y \le 1\}$ as the finest possible partition of the states of nature. Suppose, however, that we consider as states the subsets $\overline{x} = \{(x, y) : 0 \le y \le 1\}$ for each $x \in [0, 1]$. The problem that Savage discovers in this example is the following. It is possible to define small world prizes in a natural way and for preferences among small world acts to satisfy all of his axioms and, at the same time, consistently define prizes in the "grand world" consisting of the whole square S. However, it is possible for the preferences among small world acts to be consistent with the preferences among grand world acts in such a way that the probability measure determined from the small world preferences is not the marginal probability measure over the sets $\overline{x}$ induced from the grand world probability. As we will see, the problem that Savage discovers is due to using different prizes as constants in the two problems. It is not due to the small world but actually will appear in the grand world as well.

Any grand world act can be considered a small world prize. In fact, the very reason for introducing small worlds is to deal with the case in which what we count as a prize turns out to actually be worth different amounts depending on which of the subdivisions of the small world state of nature occurs. So, suppose we let

the grand world prizes be non-negative numbers and the grand world acts be all bounded measurable functions on S. The grand world probability is uniform over the square and the grand world utility is the numerical value of the prize. In order to guarantee that Savage's axioms hold in the small world, choose the small world prizes to be 0 and positive multiples of a single function h. Assuming that $U(h) = 1$, the small world probability of a set $\overline{B} = \{\overline{x} : x \in B\}$ is (from p. 89 of Savage 1954) $Q(\overline{B}) = \int_{\overline{B}} q(x)dx$, where

$$q(x) = \frac{\int_0^1 h(x,y)dy}{\int_0^1 \int_0^1 h(x,y)dydx}. \tag{22.3}$$

Unless $\int_0^1 h(x,y)dy$ is constant as a function of x, Q will not be the marginal distribution induced from the uniform distribution over S. However, even if $\int_0^1 h(x,y)dy$ is not constant, the ranking of small world acts is consistent with the ranking of grand world acts. Let $ch(\cdot,\cdot)$, considered as a small world prize, be denoted $\overline{c}$. Let $U(\overline{c}) = c$ denote the small world utility of small world prize $\overline{c}$. If $\overline{f}$ is a small world act, then for each $\overline{x}$, $\overline{f}(\overline{x}) = \overline{c}$ for some c. The expected small world utility of $\overline{f}$ is $\int_0^1 U(\overline{f}(\overline{x}))q(x)dx$. Let the grand world act f corresponding to $\overline{f}$ be defined by $f(x,y) = \overline{f}(\overline{x})h(x,y)$. It follows from (22.3) that $U(\overline{f}(\overline{x}))q(x) = \int_0^1 f(x,y)dy / \int_0^1 \int_0^1 h(x,y)dydx$. Hence, the expected small world utility of $\overline{f}$ is

$$\int_0^1 \frac{\int_0^1 f(x,y)dy}{\int_0^1 \int_0^1 h(x,y)dydx}dx,$$

which is just a constant times the grand world expected utility of f. Hence, small world acts are ranked in precisely the same order as their grand world counterparts, even though the small world probability is not consistent with the grand world probability.

We claimed that the inconsistency of the two probabilities is due to the choice of "constants" and not to the small worlds. To see this, let the grand world constants be 0 and the positive multiples of h. Then an act f in the original problem becomes an act f^* with $f^*(x,y) = f(x,y)/h(x,y)$. That is, the prize that f^* assigns to (x,y) is the number of multiples of $h(x,y)$ that $f(x,y)$ is. We define the new probability, for B a two-dimensional Borel set, $R(B) = \int_B h(x,y)dydx / \int_S h(x,y)dydx$. The expected utility of f^* is now $\int_S f^*(x,y)h(x,y)dydx / \int_S h(x,y)dydx = \int_S f(x,y)dydx / \int_S h(x,y)dydx$. This is just a constant times the original expected utility. Hence, acts are ranked in the same order by both probability-utility representations. Both representations are state-independent, but each one is relative to a different choice of constants. The constants in one representation have different utilities in different states in the other representation. Both representations satisfy Savage's axioms, however. (Note that the small world probability constructed earlier is the marginal probability associated with the grand world probability R, so that Savage's small world problem evaporates

when the definition of constant is allowed to change.) The point to remember is that the uniqueness of the probability-utility representation for a collection of preferences is relative to what counts as a constant. To use Savage's notation in the example of section "An Example", suppose that we use Yen gambles to elicit probabilities. However, instead of treating multiples of 1Y as constants, we treat multiples of gamble $f(s_1) = 100\text{Y}, f(s_2) = 125\text{Y}, f(s_3) = 150\text{Y}$ as constants. Then we will elicit the discrete uniform probability rather than the non-uniform probability.

How to Elicit Unique Probabilities and Utilities Simultaneously

There is one obvious way to avoid the confusion of the previous examples. That would be to elicit a unique probability without reference to preferences. This is the approach taken by DeGroot (1970). This approach requires that the agent have an understanding of the primitive concept "at least as likely as" in addition to the more widely understood primitive "is preferred to". Some decision theorists prefer to develop the theory solely from preference without reference to the more statistical primitive "at least as likely as". It is these latter decision theorists who need an alternative to the existing theories in order to separate probability from utility.

Karni et al. (1983) (see also Karni 1985) propose a scheme for simultaneously eliciting probability and state-dependent utility. Their scheme is essentially as follows. In addition to preferences among horse lotteries, an agent is asked to state preferences among horse lotteries under the assumption that the agent holds a particular probability distribution over the states (explicitly, they say on p. 1024, "...contingent upon a strictly positive probability distribution p' on S".) And they require the agent to compare acts with different "contingent" probabilities as well. Karni (1985) describes these (in a slightly more general setting) as *prize-state lotteries* which are functions $\hat{f}$ from $Z \times S$ to $\Re^+$ such that $\sum_{\text{all } (z,s)} \hat{f}(z,s) = 1$, and such that the probability $\hat{f}(z,s)$ for each z and s is to be understood in the same sense as the probabilities involved in the lotteries of section "State-Independent Utility". That is, the results of a prize-state lottery are determined by an auxiliary randomization. The agent is asked to imagine that the state of nature could be chosen by the randomization scheme rather than by the forces of nature. This is intended to remove the uncertainty associated with how the state of nature is determined so that a pure utility can be extracted using Axioms 1, 2, and 3 applied to a preference relation among prize-state lotteries.

For example, suppose that the agent in section "An Example" expresses a strict preference for the prize-state lottery that awards \$1 in state 2 with probability 1 ($\hat{f}(\$1, s_2) = 1$) over $\hat{g}(\$1, s_1) = 1$. This preference would not be consistent with a state-independent utility for dollar prizes, however it would be consistent with a state-independent utility in Yen prizes.

The pure utility elicited in this fashion is a function of both prizes and states, so that it is actually a state-dependent utility. So long as the preferences among prize-state lotteries are consistent with the preferences among horse lotteries, the elicited state-dependent utility can then be assumed to be the agent's utility. There will then be a unique probability such that $H_1 \preceq H_2$ if and only if the expected utility of H_1 is at most as large as the expected utility of H_2. The type of consistency that (Karni et al. 1983) require between the two sets of preferences is rather more complicated than it needs to be. The following simple consistency axiom will suffice.

Axiom 5 (Consistency). For each non-null state s and each pair $(\hat{f}_1, \hat{f}_2)$ of prize-state lotteries satisfying $\sum_{\text{all } z} \hat{f}_i(z, s) = 1$, and some pair of horse lotteries H_1 and H_2 satisfying $H_1(s_i) = H_2(s_i)$ for all $s_i \neq s$ and $H_1(s) = f_1$ and $H_2(s) = f_2$, we have $H_1 \preceq H_2$ if and only if $\hat{f}_1 \preceq \hat{f}_2$, where f_1 and f_2 are lotteries that correspond to $\hat{f}_1$ and $\hat{f}_2$ as follows: $f_i = (\hat{f}_i(z_1, s), \ldots, \hat{f}_i(z_m, s))$, $i = 1, 2$, in the notation of section "State-Independent Utility".

All that Axiom 5 says is that preferences among prize-state lotteries with all of their probabilities on the same state must be reproduced as preferences between horse-lotteries which differ only in that common state.

Theorem 2. *Suppose that there are n states of nature and m prizes. Assume that preferences among horse lotteries satisfy Axioms 1, 2, and 3. Also assume that preferences among prize-state lotteries satisfy Axioms 1, 2, and 3. Finally, assume that Axiom 5 holds. Then there exists a unique probability P over the states and a utility $U : Z \times S \to \Re$, unique up to positive affine transformation, satisfying*

1. $H_1 \preceq H_2$ *if and only if* $\sum_{i=1}^{n} P(s_i)U(H_1(s_i), s_i) \leq \sum_{i=1}^{n} P(s_i)U(H_2(s_i), s_i)$, *where, for each lottery* $L = (\alpha_1, \ldots, \alpha_m)$, $U(L, s_i)$ *stands for* $\sum_{j=1}^{m} \alpha_j U(z_j, s_i)$,
2. $\hat{f} \preceq \hat{g}$ *if and only if* $\sum_{i=1}^{n} \sum_{j=1}^{m} \hat{f}(z_j, s_i)U(z_j, s_i) \leq \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{g}(z_j, s_i)U(z_j, s_i)$.

The proof of Theorem 2 makes use of the following theorem from Fisburn (1970, p. 176):

Theorem 3 (Fishburn). *Under Axioms 1, 2, and 3, there exist real-valued functions $W_1, \ldots, W_n$ such that $H_1 \preceq H_2$ if and only if*

$$\sum_{i=1}^{n} W_i(H_1(s_i)) \leq \sum_{i=1}^{n} W_i(H_2(s_i)), \tag{22.4}$$

and the W_i that satisfy (22.4) are unique up to a similar positive linear transformations, with W_i constant if and only if s_i is null.

We provide only a sketch of the proof of Theorem 2. Let $(W_1, \ldots, W_n)$ be the state-dependent utility for horse lotteries guaranteed by Theorem 3, and let $\hat{V}$ be the utility for prize-state lotteries guaranteed by the theorem of VonNeumann and Morgenstern (1947). All we need to show is that there exist $c_1, \ldots, c_n$ and positive $a_1, \ldots, a_n$ such that for each $i = 1, \ldots, n$,

$$W_i(z) = a_i \hat{V}(z, s_i) + c_i, \text{ for all } z. \tag{22.5}$$

If (22.5) were true, then it follows directly from (22.4) that $U = \hat{V}$ would serve as the state-dependent utility and $P(s_i) = a_i / \sum_{k=1}^{n} a_k$ would be the probability. The uniqueness follows from the uniqueness of the W_i and of $\hat{V}$. To prove (22.5), let $s = s_j$ for some j and suppose that H_1, H_2, $\hat{f}_1$, $\hat{f}_2$, f_1, and f_2 are as in the statement of Axiom 5. Now, consider the set $\mathcal{H}_j$ of all horse lotteries H such that $H(s_i) = H_1(s_i)$ for all $i \neq j$. The stated preferences among this set of horse lotteries satisfies Axioms 1, 2, and 3. Hence there is a utility V_j for this set, and V_j is unique up to positive affine transformation. Clearly, W_j is such a utility, hence we will assume that $V_j = W_j$. Next, consider the set $\hat{\mathcal{H}}_j$ of all prize-state lotteries $\hat{f}$ that satisfy $\sum_{k=1}^{m} \hat{f}(z_k, s_j) = 1$. The stated preferences among elements of $\hat{\mathcal{H}}_j$ also satisfy Axioms 1, 2, and 3. Hence there is a utility $\hat{V}_j$, which is unique up to positive affine transformation. Clearly $\hat{V}$, with domain restricted to $\hat{\mathcal{H}}_j$, is such a utility, hence we will assume that $\hat{V}_j = \hat{V}$. The mapping $T_j : \mathcal{H}_j \to \hat{\mathcal{H}}_j$ defined by $T_j(H)(z, s) = 0$ for all (z, s) with $s \neq s_j$ and $T_j(H) = \alpha_i$ for $z = z_i$ and $s = s_j$, where $H(s_j) = (\alpha_1, \ldots, \alpha_m)$, is one-to-one and T_j preserves convex combination. It then follows from Axiom 5 that, for $H_1, H_2 \in \mathcal{H}_j$, $W_j(H_1) \leq W_j(H_2)$ if and only if $\hat{V}(T_j(H_1)) \leq \hat{V}(T_j(H_2))$. Since both $V_j = W_j$ and $\hat{V}_j = \hat{V}$ are unique up to positive affine transformation, we have $W_j = a_j \hat{V} + b_j$ for some positive a_j. This proves (22.5).

Discussion

The need for state-dependent utilities arises out of the possibility that what may appear to be a constant prize may not actually have the same value to an agent in all states of nature. Much of probability theory and statistical theory deals solely with probabilities and not with utilities. If probabilities are unique only relative to a specified utility, then the meaning of much of this theory is in doubt. Much of statistical decision theory makes use of utility functions of the form $U(\theta, d)$, where θ is a state of nature and d is a possible decision. The prize awarded when decision d is chosen and the state of nature is θ is not explicitly mentioned. Rather, the utility of the prize is specified without reference to the prize. Although it would appear that $U(\theta, d)$ is a state-dependent utility (as well it might be), one has swept comparisons between states "under the rug," For example, if $U(\theta, d) = -(\theta - d)^2$, one might ask how it was determined that an error of 1 when $\theta = a$ has the same utility as an error of 1 when $\theta = b$.

DeGroot (1970) avoids these problems by assuming that the concept of one event being "at least as likely as" another is understood without definition. He then proceeds to state axioms that imply the existence of a unique subjective probability distribution over states of nature. (For a discussion of attempts to derive quantitative probability from qualitative probability, see Narens 1980.) Further axioms could then be introduced that govern preference. These would then lead

to a state-dependent utility function. Axioms, such as those of Savage (1954), Von-Neumann and Morgenstern (1947), Anscombe and Aumann (1963), and deFinetti (1974), which concern only preference among acts like horse lotteries, are not sufficient to guarantee a representation of preference by a unique state-dependent utility and probability. Direct comparisons need to be made between lotteries "in a specified state of nature" and other lotteries in another specified state of nature. These are the "prize-state" lotteries introduced by Karni (1985). Assuming that preferences among prize-state lotteries are consistent with preferences among horse lotteries, a unique state-dependent utility and probability can be recovered from the preferences.

References

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of Mathematical Statistics, 34*, 199–205.

Arrow, K. J. (1974). Optimal insurance and generalized deductibles. *Scandinavian Actuarial Journal, 1*, 1–42.

deFinetti, B. (1974). *Theory of probability* (Vol. 2). New York: Wiley.

DeGroot, M. H. (1970). *Optimal statistical decisions*. New York: Wiley.

Fisburn, P. (1970). *Utility theory for decision making*. New York: Wiley.

Kadane, J. B., & Winkler, R. L. (1988). Separating probability elicitation from utilities. *Journal of the American Statistical Association, 83*, 357–363.

Karni, E. (1985). *Decision making under uncertainty*. Cambridge: Harvard University Press.

Karni, E., Schmeidler, D., & Vind, K. (1983). On state dependent preferences and subjective probabilities. *Econometrica, 51*, 1021–1031.

Narens, L. (1980). On qualitative axiomatizations for probability theory. *Journal of Philosophical Logic, 9*, 143–151.

Ramsey, F. P. (1926). Truth and probability. In H. E. Kyburg & H. E. Smokler (Eds.), *Studies in subjective probability* (pp. 23–52). Huntington: Krieger.

Rubin, H. (1987). A weak system of axioms for "rational" behavior and the non-separability of utility from prior. *Statistics and Decisions, 5*, 47–58.

Savage, L. J. (1954). *Foundations of statistics*. New York: Wiley.

Schick, F. (1986). Dutch book and money pumps. *Journal of Philosophy, 83*, 112–119.

VonNeumann, J., Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton: Princeton University Press.

Chapter 23
Causal Decision Theory

James M. Joyce and Allan Gibbard

- EXAMPLE: *Prisoner's Dilemma with Twin (PDT)*. You are caught in a standard, one-shot prisoner's dilemma (diagram next page), and the other player is your twin. You don't know for sure what Twin will do, but you know that Twin is amazingly like you psychologically. What you do, he or she too will likely do: news that you were going to rat would be good indication that Twin will rat, and news that you were going to keep mum would be a good sign that Twin will keep mum. Your sole goal is to minimize your own time in jail: Family feelings affect you not, and you care not a whit about loyalty, returning good for good, or how long Twin spends in jail. What course of action is rational for you in pursuit of your goals?

Many will find the answer easy—though they may disagree with each other on which the answer is. A standard line on the prisoner's dilemma rests on *dominance:* What you do won't affect what Twin does. Twin may rat or keep mum, but in either

In the nearly 20 years since this article was written there has been a revolution in the understanding of causal and counterfactual reasoning. This revolution had its roots in early work by Rubin (1974), Holland (1986) and Robbins (1986), which gave rise to the so-called "potential outcomes" framework. At roughly the same time the closely related "structural equations/causal graphs" approach was being developed and used to great effect by Spirtes et al. (1993), and Pearl (2000). In both treatments counterfactual reasoning plays a leading role in causal inference, just as in causal decision theory. While the core claims of this article remain true, and the basic structure of causal decision theory remains intact, these new models of provide us with far more sophisticated ways of representing and identifying causal relationships than were available and widely known when we wrote. As a result, some of our remarks about "the need for new advances in understanding of localization in relation to rational belief" have been rendered moot. Readers are encouraged to investigate these new developments, which we see as great advances.

J.M. Joyce (✉) • A. Gibbard
University of Michigan, Ann Arbor, MI, USA
e-mail: jjoyce@umich.edu; gibbard@umich.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_23

case, you yourself will do better to rat. Whichever Twin is doing, you would spend less time in jail if you were to rat than if you were to keep mum. Therefore the rational way to minimize your own time in jail is to rat.

	Mum	Rat
Mum	−1, −1	−10, 0
Rat	0, −10	−9, −9

Prisoner's Dilemma

Another line of argument leads to the opposite conclusion. Assess each act by its *auspiciousness*, by how welcome the news would be that you were about to perform it. News that you're about to rat would indicate that Twin is likewise about to rat. That's bad news; it means a long time in jail, for you as well as for Twin. News that you're about to keep mum, on the other hand, would be good news: It indicates that Twin is likewise about to keep mum, and your both keeping mum will mean a short time in jail. Keeping mum, then, is the *auspicious* act, and so—in terms of your selfish goals—you achieve best prospects by keeping mum.[1]

The two lines of reasoning, then, lead to opposite conclusions. One or the other, to be sure, may strike a reader as obviously wrong. Still, if one of them is cogent and the other not, decision theory should tell us why. Standard theories haven't spoken, though, with one voice on this matter. Savage himself (1972) was mostly silent on issues that would decide between the two lines: his system could be read in more than one way, and the few pertinent remarks he left us point in opposing directions. Various other decision-theoretic systems do have implications for this matter. Some imply that the argument from auspiciousness is correct: the principle of dominance, these systems entail, doesn't properly apply to a case like PDT. Taking the other side, a group called—perhaps somewhat misleadingly—*causal* decision theorists have formulated systems according to which the principle of dominance does apply to this case, and the rational thing to do is to rat.

"Causal" theorists maintain that decision theory requires a notion of causal dependency, explicit or implicit. Otherwise, they say, the theory will yield the wrong prescription for cases like PDT. We touch below on how causal notions might be made explicit for decision theorists' purposes, and how causality might be vindicated as empirically respectable. Auspiciousness theorists—or *evidential*

[1] An interesting model of Prisoner's Dilemma with a twin can be found in Howard (1988). Howard, who endorses a version of the auspiciousness argument, shows how to write a Basic program for playing the game which is capable of recognizing and cooperating with programs that are copies of itself.

decision theorists, as they are called in the literature—have no need for causal terms in their theory: they manage everything with standard subjective probabilities and conditional probabilities. Some evidential theorists deny that their theory, properly construed or developed, really does say not to rat in PDT. They deny, then, that causal notions must be introduced into decision theory, even if causal theorists are right about what to do in this case. We touch below on debates between "causal" theorists and this camp of "evidential" theorists, but mostly stick with the "causal" theory, explaining it and examining its potentialities.[2]

Cases with the structure of PDT can't be rare. The prisoner's dilemma itself is a parable, but economics, politics, war, and the like will be full of cases where one's own acts suggest how others are acting. Consider, for instance, a sophisticated speculator playing a market. Mustn't he reasonably take himself to model other sophisticated players? Why should he be unique? A rational agent interacting with others must escape the hubris of thinking that only he is smart and insightful—but then he'll have to take himself as a likely model for the schemings and reasonings of others. In such cases, different versions of decision theory may prescribe incompatible actions.

Dominance and Expected Utility: Two Versions

Savage (1972) encoded decision problems as matrices, with columns indicating "states of the world". For a Savage matrix with states $S_1, \ldots, S_n$, the expected utility of an act A is calculated by the Savage formula

$$\mathcal{V}(A) = \sum_{i=1}^{n} \rho(S_i)u(A, S_i), \tag{23.1}$$

where $u(A, S_i)$ is the utility of act A for state S_i and $\rho(S_i)$ is the subjective probability of S_i. From (23.1) follows a principle which we'll call the *Unqualified Principle of Dominance*, or *UPD:*

- UPD: If for each S_i, $u(A, S_i) > u(B, S_i)$, then $\mathcal{V}(A) > \mathcal{V}(B)$.

Which Savage matrix correctly represents a problem, though, must be decided with care, as is shown by a spoof due to Jeffrey (1967), p. 8:

[2]Nozick (1969) introduced PDT and other cases of this kind, focusing his discussion on Newcomb's problem, which he credits to physicist William Newcomb. He makes many of the points that causal theorists have come to accept, but recognizes only one kind of expected utility, the one we are calling auspiciousness. Stalnaker originated causal expected utility in a 1972 letter published only much later (Stalnaker 1981). Gibbard and Harper (1978) proselytize Stalnaker's proposal, and Lewis (1981) gives an alternative formulation which we discuss below. Gibbard and Harper (1978) and Lewis (1979b) also discuss PDT along with Newcomb's problem.

- EXAMPLE: *Better Red Than Dead (BRD)*. I'm an old-time American cold warrior with scruples, deciding whether or not my country is to disarm unilaterally. I construct a matrix as follows: My two possible states are that the Soviets invade and that they don't. In case they invade, better red than dead; in case they don't, better rich than poor. In either case, unilateral disarmament beats armament, and so by dominance, I conclude, it is rational to disarm.

Now whether or not unilateral disarmament would be rational all told, this argument can't be right. As the proponent of deterrence will point out, unilateral disarmament may decrease the likelihood of the Soviets' invading, and a scenario in which they don't invade is better than one in which they do. The argument from "dominance" treats these considerations as irrelevant—even if the Soviets are sure to invade if we disarm and to hold back if we arm.

Savage's states, then, must be act-independent: They must obtain or not independently of what the agent does. How, then, shall we construe this requirement? The first answer was developed, independently, by Jeffrey (1967) and by Luce and Krantz (1971): For dominance correctly to apply, they say, the states must be stochastically (or probabilistically) independent of acts. Where acts $A_1 \ldots A_m$ are open to the agent and $\rho(S/A_j)$ is the standard conditional probability of S given A_j,[3] S is stochastically act-independent iff

$$\rho(S/A_1) = \rho(S/A_2) = \cdots = \rho(S/A_m) = \rho(S) \tag{23.2}$$

The probabilities in question are subjective, or at any rate epistemic. Formula (23.2), then, means roughly that learning what one's going to do won't affect one's credence in S. One's act won't be evidence for whether S obtains. Requirement (23.2), then, is that any state S be *evidentially* act-independent. Theories that entail that (23.2) is what's needed for the principle of dominance to hold good we can thus call *evidential* theories of decision.

As applied to the prisoner's dilemma with your twin, evidential decision theory must treat the dominance argument as bad. Twin's act fails to be independent of yours evidentially. Let proposition C_t be that Twin cooperates, let C_y be that you cooperate, and let D_y be that you defect. C_t is not evidentially independent of what you do: $\rho(C_t/C_y)$ is high whereas $\rho(D_t/C_y)$ is low, since your cooperating is good evidence that she is cooperating, whereas your defecting is good evidence that she is defecting. Condition (23.2), then, won't hold for this case.

Evidential decision theory holds that the Savage formula (23.1) applies when condition (23.2) holds—that is, when the states of the matrix are evidentially act-independent. Requirements of act-independence, though, could be dropped if we changed formula (23.1) for computing expected utility. Use as weights not the probability of each state, as in (23.1), but its conditional probability—its probability conditional on the act's being performed. In this more general formulation, we have

[3] More precisely, A_i is the *proposition* that one performs a particular one of the alternative acts open to one in one's circumstances. We reserve the notation $\rho(S|A_j)$ for a more general use later in this chapter.

$$\mathcal{V}(A) = \sum_{i=1}^{n} \rho(S_i/A)u(A, S_i). \tag{23.3}$$

The Savage formula (23.1) is then a special case of (23.3), for conditions of evidential act-independence.[4] Since UPD follows from (23.1), it follows from (23.3) plus condition (23.2) of evidential act-independence. Evidential decision theory, then, has (23.3) as its general formula for expected utility. Its version of the principle of dominance is UPD qualified by condition (23.2) of evidential act-independence. In general, it recommends using "auspiciousness" to guide choices: a rational agent should select an act whose performance would be best news for him—roughly, an act that provides best evidence for thinking desirable outcomes will obtain.

Evidential theory has the advantage of avoiding philosophically suspect talk of causality: Its general formula (23.3) sticks to mathematical operations on conditional probabilities, and likewise, its requirement (23.2) of evidential act-independence—its condition, that is, for the Savage formula (23.1) to apply to a matrix—is couched in terms of conditional probabilities.

The causal theorist, in contrast, maintains that to apply a principle of dominance correctly, one can't avoid judgments of causality. One must form degrees of belief as to the causal structure of the world; one must have views on what is causally independent of what. Belief in Twin's causal isolation is a case in point. Dominance applies to PDT, their contention is, because you and your twin are causally isolated—and you know it. What you do, you know, will in no way affect what twin does. The argument, then, invokes a causal notion: the notion of what will and what won't causally *affect* what else. Causal decision theory then recommends using causal efficacy to guide choices: It holds, roughly, that a rational agent should select an act whose performance would be likely to bring about desirable results.

The causal decision theorist's requirement on a state S of a Savage matrix, then, is the following: The agent must accept that nothing he can do would causally affect whether or not S obtains. For each act A open to him, he must be certain that S's obtaining would not be causally affected by his doing A.

Can the causal theorist find a formula for expected utility that dispenses with this requirement of believed causal act-independence? A way to do so was proposed by Stalnaker (1968); see also Gibbard and Harper (1978). It requires a special conditional connective, which we'll render '$\Box\!\!\rightarrow$'. Read '$A \Box\!\!\rightarrow B$' as saying, "If A obtained then B would." In other words, either A's obtaining would cause B to obtain, or B obtains independently (causally) of whether or not A obtains. Then to say that S is causally independent of which act $A_1 \ldots A_n$ one performs is to say this: Either S would hold whatever one did, or whatever one did S would fail to hold. In other words, for every act A_i, we have $A_i \Box\!\!\rightarrow S$ iff S. We can now generalize the Savage formula (23.1) for the causal theorist's kind of expected utility. Use as weights, now, the probabilities of conditionals $\rho(A \Box\!\!\rightarrow Si)$, as follows:

[4]Formula (11.1) is introduced by Jeffrey (1967) and by Luce and Krantz (1971).

$$\mathcal{U}(A) = \sum_{i=1}^{n} \rho(A \,\square\!\!\rightarrow S_i) u(A, S_i) \tag{23.4}$$

Call this $\mathcal{U}(A)$ the *instrumental expected utility* of act A. The Savage formula

$$\mathcal{U}(A) = \sum_{i=1}^{n} \rho(S_i) u(A, S_i), \tag{23.5}$$

is then (23.4) for the special case where the following condition holds:

$$\text{For each } S_i, \rho\,(A \,\square\!\!\rightarrow S_i) = \rho\,(S_i)\,. \tag{23.6}$$

A sufficient condition for this to hold is that, with probability one, S_i is causally independent of A—in other words, that

$$\text{For each } S_i, \rho\,([A \,\square\!\!\rightarrow S_i] \leftrightarrow S_i) = 1. \tag{23.7}$$

Note that for the prisoner's dilemma with twin, condition (23.7) does hold. Twin, you know, is causally isolated from you. You know, then, that whether Twin would defect if you were to defect is just a matter of whether twin is going to defect anyway. In other words, you know that $D_y \,\square\!\!\rightarrow D_t$ holds iff D_t holds, and so for you, $\rho([D_y \,\square\!\!\rightarrow D_t] \leftrightarrow D_t) = 1$. This is an instance of (23.7), and similar informal arguments establish the other needed instances of (23.7) for the case.

In short, then, causal decision theory can be formulated taking formula (23.4) for instrumental expected utility as basic. It is instrumental expected utility as given by (23.4), the causal theorist claims, that is to guide choice. The Savage formula is then a special case of (23.4), for conditions of known causal act-independence—where (23.7) holds, so that for each state S_i, $\rho(A \,\square\!\!\rightarrow S_i) = \rho(S_i)$. The Unqualified Principle of Dominance for $\mathcal{U}$ is

- UPD: If for each S_i, $u(A, S_i) > u(B, S_i)$, then $\mathcal{U}(A) > \mathcal{U}(B)$.

Causal decision theory, in this formulation, has (23.4) as its general formula for the instrumental expected utility that is to guide choice, and its own version of the principle of dominance: UPD qualified by condition (23.7) of known causal act-independence.

Evidential and causal decision theorists, in short, accept different general formulas for the expected utility that is to guide choice, and consequently, they accept different conditions for the Savage formula to apply, and different principles of dominance. Causal theory—in the formulation we've been expounding—adopts (23.4) as its formula for expected utility, whereas evidential theory adopts (23.3). Causal theory, in other words, weighs the values of outcomes by the probabilities of the relevant conditionals, $\rho(A \,\square\!\!\rightarrow S_i)$, whereas evidential theory weighs them by the relevant conditional probabilities $\rho(S_i/A)$. Different conditions, then, suffice, according to the two theories, for the Savage formula correctly to apply to a matrix,

and consequently for UPD to apply. That makes for distinct principles of dominance: For the causal theorist, UPD qualified by condition (23.7) of known causal act-independence, and for the evidential theorist, UPD qualified by condition (23.2) of evidential act-independence.

Conditionals and Their Probabilities

What, then, is the contrast on which all this hinges: the contrast between the probability $\rho(A \mathbin{\square\!\!\rightarrow} S)$ of a conditional $A \mathbin{\square\!\!\rightarrow} S$ and the corresponding conditional probability $\rho(S/A)$? Where probability measure ρ gives your *credences*—your degrees of belief—the conditional probability $\rho(S/A)$ is the degree to which you'd believe S if you learned A and nothing else. In the prisoner's dilemma with your twin, then, $\rho(D_t/D_y)$ measures how much you'd expect twin to rat on learning that you yourself were about to rat. If $\rho(D_t/D_y) \neq \rho(D_t/C_y)$, that doesn't mean that D_t is in any way causally dependent on whether D_y or C_y obtains. It just means that your act is somehow *diagnostic* of Twin's. Correlation is not causation. Probability $\rho(D_y \mathbin{\square\!\!\rightarrow} D_t)$, on the other hand, is the degree to which you believe that if you were to defect, then Twin would. There are two circumstances in which this would obtain: Either Twin is about to defect whatever you do, or your defecting would cause Twin to defect. To the degree to which $\rho(D_y \mathbin{\square\!\!\rightarrow} D_t) > \rho(D_y \mathbin{\square\!\!\rightarrow} C_t)$, you give some credence to the proposition $[D_y \mathbin{\square\!\!\rightarrow} D_t]\&\neg[D_y \mathbin{\square\!\!\rightarrow} C_t]$, that twin would defect if you did, but not if you cooperated. This is credence in the proposition that your act will make a causal difference.

In daily life, we guide ourselves by judgments that seem to be conditional: What would happen if we did one thing, or did another? What would be the effects of the various alternatives we contemplate? We make judgments on these matters and cope with our uncertainties. Classic formulations of decision theory did not explicitly formalize such notions: notions of causal effects or dependency, or of "what would happen if". In Ramsey's and Savage's versions, causal dependency may be implicit in the representational apparatus, but this formal apparatus is open to interpretation. Other theorists had hoped that whatever causal or "would" beliefs are involved in rational decisions could be captured in the structure of an agent's conditional probabilities for non-causal propositions or events.

This last maneuver might have great advantages if it worked, but causal decision theorists argue that it doesn't. Causal or "would" notions must somehow be introduced into decision theory, they claim, if the structure of decision is to be elucidated by the theory. The introduction can be explicit, as in the general forumula (23.4) for $\mathcal{U}$ above, or it can be in the glosses we give—say, in interpretations of the Savage formula (23.1) or (23.5). If causal theorists are right, then, the theory of subjective conditional probability won't give us all we need for describing the beliefs relevant to decision. We'll need some way of displaying such beliefs and theorizing about them.

Causal theorists have differed, though, on how causal beliefs are best represented. So far, we've spoken in Stalnaker's terms, but we need to say more on what his treatment consists in, and what some of the alternatives might be for representing causal decision theory.

First, some terminology. Savage spoke of "states" and "events", and distinguished these from "acts" or "strategies". The philosophers who developed causal decision theory often spoke of "propositions", and include as propositions not only Savage's "events" and "states", but also acts and strategies. That is to say, propositions can characterize not only what happens independently of the agent, but also what the agent does—or even what he *would* do in various eventualities. A proposition can say that I perform act a or adopt strategy s. Such propositions can be objects of belief and of desire, and so can be assigned credences (subjective probabilities) and utilities.

Let A, then, be the proposition that I perform act a. Stalnaker constructs a conditional proposition $A \mathbin{\square\!\!\rightarrow} B$, which we read as "If I did A, then B would obtain." How does such a conditional proposition work? Much as Savage treats an event as a set of states, so Stalnaker treats a proposition as a set of *possible worlds* or maximally specific ways things might have been. Abstractly, the connective '$\mathbin{\square\!\!\rightarrow}$' is a two-place propositional function: To each pair of propositions it assigns a proposition.

Stalnaker hoped originally that this conditional function $\mathbin{\square\!\!\rightarrow}$ could be defined so that the probability of a conditional is always the corresponding conditional probability: So that whenever $\rho(C/A)$ is defined, $\rho(A \mathbin{\square\!\!\rightarrow} C) = \rho(C/A)$. Lewis (1976) proved that—with trivial exceptions—no such equality will survive conditionalization. Read ρ_A, in what follows, as probability measure ρ conditioned on A, so that by definition, $\rho_A(C) = \rho(C/A)$. What Lewis showed impossible is this: that for all propositions A, C and B for which $\rho(A\&B) > 0$, one has $\rho_B(A \mathbin{\square\!\!\rightarrow} C) = \rho_B(C/A)$. For if this did obtain, then one would have both $\rho_C(A \mathbin{\square\!\!\rightarrow} C) = \rho_C(\mathrm{C/A})$ and $\rho_{\neg C}(A \mathrm{\ W\ } C) = \rho_{\neg C}(C/A)$. But then

$$\begin{aligned}
\rho(A \mathbin{\square\!\!\rightarrow} C) &= \rho_C(A \mathbin{\square\!\!\rightarrow} C)\,\rho(C) + \rho_{\neg C}(A \mathbin{\square\!\!\rightarrow} C)\,\rho(\neg C) \\
&= \rho_C(C/A)\,\rho(C) + \rho_{\neg C}(C/A)\,\rho(\neg C) \\
&= 1 \cdot \rho(C) + 0 \cdot \rho(\neg C) \\
&= \rho(C)
\end{aligned}$$

We'd have $\rho(A \mathbin{\square\!\!\rightarrow} C) = \rho(C/A)$, then, at most when $\rho(C) = \rho(C/A)$. No such equality can survive conditionalization on an arbitrary proposition.

How, then, should we interpret the probability $\rho(A \mathbin{\square\!\!\rightarrow} C)$ of a conditional proposition $A \mathbin{\square\!\!\rightarrow} C$, if it is not in general the conditional probability $\rho(C/A)$. Many languages contrast two forms of conditionals, with pairs like this one[5]:

[5] Adams (1975) examines pairs like this.

If Shakespeare didn′t write *Hamlet*, someone else did. (23.8)

If Shakespeare hadn′t written *Hamlet*, someone else would have. (23.9)

Conditionals like (23.8) are often called *indicative*, and conditionals like (23.9) *subjunctive* or *counterfactual*. Now indicative conditional (23.8) seems epistemic: To evaluate it, you might take on, hypothetically, news that Shakespeare didn't write Hamlet. Don't change anything you now firmly accept, except as you would if this news were now to arrive. See, then, if given this news, you think that someone else did write Hamlet. You will, because you are so firmly convinced that Hamlet was written by someone, whether or not the writer was Shakespeare. The rule for this can be put in terms of a thinker's subjecive probabilities—or her *credences*, as we shall say: indicative conditional (23.8) is acceptable to anyone with a sufficiently high conditional credence $\rho(E/D)$ that someone else wrote Hamlet given that Shakespeare didn't. Subjunctive conditional (23.9) works differently: If you believe that Shakespeare did write Hamlet, you will find (23.9) incredible. You'll accept (23.8), but have near zero credence in (23.9). Your conditional credence $\rho(E/D)$ in someone else's having written Hamlet given that Shakespeare didn't will be high, but your credence $\rho(D \,\square\!\!\rightarrow E)$ in the subjunctive conditional proposition (23.9) will be near zero. Here, then, is a case where one's credence in a conditional proposition (23.9) diverges from one's corresponding conditional credence. Speaking in terms of the subjective "probabilities" that we have been calling credences, we can put the matter like this: the probability $\rho(D \,\square\!\!\rightarrow E)$ of a subjunctive conditional may differ from the corresponding conditional probability $\rho(E/D)$.

The reason for this difference lies in the meaning the $\square\!\!\rightarrow$ operator. Stalnaker (1968) puts his account of conditionals in terms of alternative "possible worlds". A world is much like a "state" in Savage's framework (except that it will include a strategy one might adopt and its consequences). Think of a possible world as a maximally specific way things might have been, or a maximally specific consistent proposition that fully describes a way things might have been. Now to say what the proposition $A \,\square\!\!\rightarrow C$ is, we have to say what conditions must obtain for it to be true. There is no difficulty when the antecedent A obtains, for then, clearly, $A \,\square\!\!\rightarrow C$ holds true if and only if C obtains. The puzzle is for cases where A is false. In those situations, Stalnaker proposes that we imagine the possible world w^A in which A is true, and that otherwise is most *similar* to our actual world in relevant respects. $A \,\square\!\!\rightarrow C$, then, holds true iff C holds in this world w^A. Stalnaker and Thomason offered a rigorous semantics and representation theorem for this explication. Stalnaker's distinctive axioms are these[6]:

- *Intermediate strength:* If A necessitates B, then $A \,\square\!\!\rightarrow B$, and if $A \,\square\!\!\rightarrow B$, then $\neg(A\&\neg B)$
- *Conditional non-contradiction:* For possible A, $\neg[(A \,\square\!\!\rightarrow B)\&(A \,\square\!\!\rightarrow \neg B)]$.

[6]Stalnaker (1968) p. 106, Stalnaker and Thomason (1970), slightly modified.

- *Distribution:* If $A \mathrel{\square\!\!\rightarrow} (B \vee C)$, *then* $(A \mathrel{\square\!\!\rightarrow} B) \vee (A \mathrel{\square\!\!\rightarrow} C)$.
- *Suppositional equivalence:* If $(A \mathrel{\square\!\!\rightarrow} B)$ and $(B \mathrel{\square\!\!\rightarrow} A)$, then $(A \mathrel{\square\!\!\rightarrow} C)$ iff $(B \mathrel{\square\!\!\rightarrow} C)$.

All this leaves mysterious the notion of *relevant similarity* invoked by Stalnaker's account. Formal axioms are easy: Stalnaker speaks of a *selection function* f which assigns a world $f(A, w) = w^A$ to each proposition A that has the possibility of being true. A compelling logic for $\mathrel{\square\!\!\rightarrow}$ can be developed on this basis.

How, though, do we interpret the notion of "relevant similarity" when applying this formal apparatus to real-life decision problems? Intuitive overall likeness of worlds won't do. Nixon, imagine, had in the Oval Office a red button to launch the missiles.[7] In his despair he considered pushing it, but drew back. We can say, "If Nixon had pushed the button, nuclear holocaust would have ensued." This is true, we would judge, if the apparatus was in working order and without safeguards. Of the worlds in which Nixon pushes the button, though, the one most similar overall to the actual world would be not one in which all was destroyed, but one in which the apparatus malfunctioned—a wire became temporarily nonconducting, say. After all, a world in which the missiles were launched would surely have a future radically different from the actual world. Little in any city would look the same.

Clearly, then, we cannot look to overall similarity of worlds to cash out the type of "relevant" similarity needed in the evaluation of subjunctive conditionals. Rather, we want to know what would have ensued from initial conditions that were much like those that actually obtained, but differed in some slight respect in Nixon's decision-making apparatus—differed in such a way that by natural laws, the outgrowth of those modified initial conditions would have been nuclear holocaust. Thus, the rough idea is that one evaluates it $A \mathrel{\square\!\!\rightarrow} C$ by considering a world in which A obtains that is as much like the actual world as possible both with regard to particular facts about the past as well as general facts about what might follow causally from what in the future. Lewis (1979a) attempts a general account of the kind of "relevant similarity" that fits our causal judgments, and derives from it an account of the regularities that give time its direction. As decision theorists, though, we need not be concerned whether such lofty philosophical ambitions can be fulfilled. We need only understand that where w_0 is the actual world, the value $f(A, w_0)$ of the Stalnaker selection function is the world as it would be if A obtained. It is that world, with all its ensuing history. In many situations, it will be clear that agents do have subjective probabilities for what that world would be like, and so the application of Stalnaker's apparatus to an agent's decision situation will be clear enough.[8]

Stalnaker's framework allows us to be more precise about how probabilities of subjunctive conditionals differ from ordinary conditionals probabilities. It is useful

[7] Fine (1975) gives this example to make roughly this point.

[8] Shin (1991a), for instance, devises a metric that seems suitable for simple games such as "Chicken".

to think of both $\rho(A \mathrel{\square\!\!\rightarrow} C)$ and $\rho(C/\!/A)$ as capturing a sense of *minimal* belief revision. $\rho(C/\!/A)$, as we have seen, is the credence that an agent with prior ρ should assign C if she gets pure and certain news of A's truth. Thus, the function $\rho_A = \rho(\bullet /\!/ A)$ describes the outcome of a belief revision process in which an agent learns that A. This revision is minimal in the sense that it changes ρ so as to make A certain without thereby altering any ratios of the form $\rho(X\&A) : \rho(Y\&A)$. In terms of possible worlds, ρ_A is obtained from ρ by setting the credence of $\neg A$ equal to zero, and spreading the residual probability uniformly over the worlds in which A obtains—thus leaving undisturbed any evidential relationships that might obtain among propositions that entail A. The function $\rho^A = \rho(A \mathrel{\square\!\!\rightarrow} \bullet)$ defines a rather different sort of minimal belief revision, *imaging* (Lewis 1976, 1981) or better, *imagining*. Instead of distributing the probability of $\neg A$ uniformly over the A-worlds, spread it with an eye to relevant similarities among worlds. Transfer the probability of each world w in which A is false to w^A, the A-world most similar to w, adding this probability to the probability w^A has already.

This whole treatment, though, rests on an assumption that is open to doubt: that there always exists a unique "most similar" world w^A. Perhaps there's no one definite way the world would be were I, say, now to flip this coin. The world might be indeterministic, or the supposition that I now flip the coin might be indeterminate—in that the exact force and manner in which I'd be flipping the coin isn't specified and isn't even under my control. It may then be the case neither that definitely were I to flip the coin it would land heads, nor that definitely were I to flip the coin it would not land heads. The law of *conditional excluded middle* will be violated:

$$(F \mathrel{\square\!\!\rightarrow} H) \vee (F \mathrel{\square\!\!\rightarrow} \neg H) \tag{23.10}$$

Conditional excluded middle obtains in Stalnaker's model. There are models and logics for conditionals in which it does not obtain.[9] In a case like this, however, the right weight to use for decision is clearly not the probability $\rho(F \mathrel{\square\!\!\rightarrow} H)$ of a conditional proposition $F \mathrel{\square\!\!\rightarrow} H$. If I'm convinced that neither $F \mathrel{\square\!\!\rightarrow} H$ nor $F \mathrel{\square\!\!\rightarrow} \neg H$ obtains, then my subjective probability for each is zero. The weight to use if I'm betting on the coin, though, should normally be one-half.

A number of ways have been proposed to handle cases like these. One is to think that with coins and the like, there's a kind of conditional chance that isn't merely subjective: the chance with which the coin would land heads *were* I to flip it. Write this $\pi_F(H)$. Then use as one's decision weight the following: one's *subjectively expected value for* this *objective conditional chance*. Suppose you are convinced that the coin is loaded, but don't know which way: You think that the coin is loaded either .6 toward heads or .6 toward tails, and have subjective probability of .5 for each of these possibilities:

[9]Lewis (1973) constructs a system in which worlds may tie for most similar, or it may be that for every A-world, there is an A-world that is more similar. He thus denies Conditional Excluded Middle: It fails, for instance, when two A-worlds tie for most similar to the actual world, one a C-world and the other a $\neg C$-world.

$$\rho\left(\pi_F(H) = .6\right) = .5 \qquad \text{and} \qquad \rho\left(\pi_F(H) = .4\right) = .5. \tag{23.11}$$

Your subjectively expected value for $\pi_F(H)$, then, will be the average of .6 and .4. Call this appropriate decision weight $\varepsilon_F(H)$. We can express this weighted averaging in measure theoretic terms, so that in general,

$$\varepsilon_A(C) = \int_0^1 x \cdot \rho\left(\pi_A(C) \in dx\right). \tag{23.12}$$

$\varepsilon_A(C)$ thus measures one's subjective expectation of C's obtaining were A to occur: the sum of (i) the degree to which A's obtaining would tend to bring it about that C obtained, plus (ii) the degree to which C would tend to hold whether or not A obtained.

We can now write formulas for $\mathcal{U}$ using $\varepsilon_A(C)$ where we had previously used $\rho(A \mathbin{\square\!\!\rightarrow} C)$. Formula (23.4) above for instrumental expected utility now becomes

$$\mathcal{U}(A) = \sum_{i=1}^{n} \varepsilon_A(S_i) u(A, S_i) \tag{23.13}$$

Lewis (1981) gives an alternative formulation of causal decision theory, which is equivalent to the Stalnaker formulation we've been presenting whenever the Stalnaker framework applies as intended. He speaks of *dependency hypotheses:* complete hypotheses concerning what depends on what and how. Which dependency hypothesis obtains is causally independent of what the agent does, and so a Lewis dependency hypothesis can serve as a "state" in the Savage framework. He allows dependency hypotheses to contain an element of objective chance.[10]

Can an empiricist believe in such things as objective conditional chance or objective dependencies? This is a hotly debated topic, mostly beyond the scope of this article. Some philosophers think that objective dependency can be defined or characterized, somehow, in purely non-causal terms. Others doubt that any such direct characterization is possible, but think that a more indirect strategy may be available: Characterize a thinker's *beliefs* in causal dependencies—or his *degrees* of belief, his subjective probabilities, or as we have been calling them, his *credences*. His credences in causal propositions, this contention is, can be cashed out fully in terms of complex features of his credences in non-causal propositions—propositions that don't involve causal notions.

We ourselves would argue that causal propositions are genuine propositions of their own kind, basic to thinking about the world. They can't be fully explained in other terms, but they can be vindicated. We can be just as much empiricists about

[10]Skyrms (1980) offers another formulation, invoking a distinction between factors that are within the agent's control and factors that aren't. Lewis (1981) discusses both Skyrms and unpublished work of Jordan Howard Sobel, and Skyrms (Skyrms 1984), 105–6, compares his formulation with those of Lewis (1981) and Stalnaker (1981).

causes as we can about other features of the layout of the world. A rational thinker forms his credences in causal propositions in much the same Bayesian way he does for any other matter: He updates his subjective probabilities by conditionalizing on new experience. He starts with reasonable prior credences, and updates them. Subjective probability theorists like de Finetti long ago explained how, for non-causal propositions, updating produces convergence. The story depends on surprisingly weak conditions placed on the thinker's prior credence measure. The same kind of story, we suspect, could be told for credences in objective chance and objective dependence.

Lewis (1980) has told a story of this kind for credence in objective chance. His story rests on what he labels the "Principal Principle", a condition which characterizes reasonable credences in objective chances. Take a reasonable credence measure ρ, and a proposition about something that hasn't yet eventuated—say, that the coin I'm about to flip will land heads. Let me conditionalize his credences on the proposition that as of now, the objective probability of this coin's landing heads is .6. Then his resulting conditional credence in the coin's landing heads, the principle says, will likewise be .6. Many features of reasonable credence in objective chance follow from from this principle. From a condition on reasonable prior credences in objective chance follows an account of how one can learn about them from experience.

A like project for objective dependency hasn't been carried through, so far as we know, but the same broad approach would seem promising.[11] In the meantime, there is much lore as to how experience can lead us to causal conclusions—and even render any denial of a causal depencency wildly implausible. The dependence of cancer on smoking is a case in point. Correlation is not causation, we all know, and a conditional probability is not a degree of causal dependence. (In the notation introduced above, the point is that $\rho(C/A)$ need not be $\varepsilon_A(C)$, the subjectively expected value of the objective chance of C were one to do a.) Still, correlations, examined with sophistication, can *evidence* causation. A chief way of checking is to "screen off" likely common causes: A correlation between smoking and cancer might arise, say, because the social pressures that lead to smoking tend also to lead to drinking, and drinking tends to cause cancer. A statistician will check this possibility by separating out the correlation between smoking and cancer among drinkers on the one hand, and among non-drinkers on the other. More generally, the technique is this: A correlation between A and C, imagine, is suspected of being spurious—suspected *not* to arise from a causal influence of A on C or *vice versa*. Let F be a suspected common cause of A and C that might account for their correlation. Then see if the correlation disappears with F held constant. Econometricians elaborate such devices to uncover causal influences in an economy. The methodological literature on gleaning causal conclusions from experience includes classic articles by Herbert Simon (see Simon (1957), chs. 1–3).

[11] The work of Spirtes et al. (1993) and Pearl (2000) goes a long way toward realizing this goal.

Screening off is not a sure test of causality. A correlation might disappear with another factor held constant, not because neither factor depends causally on the other, but because the causal dependency is exactly counterbalanced by a contrary influence by a third factor.[12] Such a non-correlation might be robust, holding reliably with large sample sizes. But it will also be a coincidence: opposing tendencies may happen to cancel out, but we can expect such cases to be rare. Lack of correlation after screening off is *evidence* of lack of causal influence, but doesn't *constitute* lack of causal influence.

When controlled experiments can be done, in contrast, reasonable credences in a degree of objective dependency can be brought to converge without limit as sample size increases. Subjects are assigned to conditions in a way that we all agree has no influence on the outcome: by means of a chance device, say, or a table of pseudo-random numbers. Observed correlations then evidence causal dependence to whatever degree we can be confident that the correlation is no statistical fluke. With the *right* kind of partition, then, screening off does yield a reliable test of causality. But what makes a partition suitable for this purpose, we would claim, must be specified in terms that somehow invoke causality—in terms, for instance, of known causal independence.

How we can use evidence to support causal conclusions needs study. Standard statistical literature is strangely silent on questions of causation, however much the goals of statistical techniques may be to test and support causal findings. If we are right, then one class of treatments of causality will fail: namely, attempts to characterize causal beliefs in terms of the subjective probabilities and the like of non-causal propositions. Correct treatments must take causality as somehow basic. A constellation of relations—cause, chance, dependence, influence, laws of nature, what would happen if, what might likely happen if—are interrelated and may be intercharacterizable, but they resist being characterized purely from outside the constellation. Our hope should be that we can show how the right kind of evidence lets us proceed systematically from causal truisms to non-obvious causal conclusions. Fortunately, much of decision and game theory is already formulated in terms that are causal, implicitly at least, or that can be read or interpreted as causal. (When games are presented in normal form, for instance, it may be understood that no player's choice of strategies depends causally on the choice of any other.) A chief aim of causal decision theory is to make the role of causal beliefs in decision and game theory explicit.

Ratificationism

While some proponents of auspiciousness maximization have taken the heroic course and tried to argue that cooperating in Prisoner's Dilemma with Twin is rational,[13] most now concede that only defection makes sense. Nevertheless,

[12] Gibbard and Harper (1978), 140–2, construct an example of such a case.

[13] See, for example, Horgan (1981).

the promise of an analysis of rational choice free from causal or counterfactual entanglements remains appealing. A number of writers have therefore sought to modify the evidential theory so that it endorses the non-cooperative solution in PDT and yet does not appeal any unreconstructed causal judgements. The hope is that one will be able to find statistical techniques of the sort discussed toward the end of the last section to distinguish causal relationships from suprious correlations, and then employ these techniques to formulate a decision theory that can be sensitive to causal considerations without making explicit use of causal or subjunctive notions. The most influential of these attempts are found in Eells (1982) and Jeffrey (1983). Since the two approaches are similar, we focus on Jeffrey.

As we have seen, statisticians sometimes use "screening off" techniques to detect the effects of a common cause. Jeffrey's strategy is based on the insight that an agent's ability to anticipate her own decisions typically screens off any purely evidential import that her actions might possess. Prior to performing an act she will generally come to realize that she has decided on it. The act itself then ceases to be a piece of evidence for her since she has already discounted it. Letting $\triangle^A$ denote the decision to do a, we can put Jeffrey's claim like this:

- *Screening*. The decision to perform A screens off any purely evidential correlations between acts and states of the world, in the sense that $\rho(S/B\&\triangle^A) = \rho(S/\triangle^A)$ for all acts B and states S.

To ensure that these conditional credences are well-defined, Jeffrey must assume that the agent always assigns some positive probability to the prospect that she will fail to carry out a decision—due to a "trembling hand," a lack of nerve, or other factors beyond her control—so that $\rho(B\&\triangle^A)$ is non-zero for all acts B.

To see what happens when screening is introduced into PDT, imagine that during the course of your deliberations but prior to performing any act, you become certain that you will decide to cooperate, so that your credence in Δ_y^C moves to one. Since you are likely to carry out whatever decision you make, your probability for C_y also moves close to one, which gives you strong grounds for thinking your twin will cooperate. Indeed, if Screening obtains, you will have strong evidence for thinking that Twin is about to cooperate, no matter what news you get as to what you yourself are about do, because $\rho\left(C_t/C_y\&\Delta_y^C\right) = \rho\left(C_t/D_y\&\Delta_y^C\right) = \rho\left(C_t/\Delta_y^C\right) \approx 1$. Condition (23.2) is thus satisfied. You can then correctly apply the evidential dominance principle to conclude that defection is your most auspicious option. In this way, Screening ensures that if you are certain that you will eventually decide to cooperate, then you will assign defecting a higher auspiciousness than cooperating. On the other hand, if you are certain that you will decide to defect, you then have strong reason to suspect that twin is about to defect, whatever you learn that you yourself are about to do—and again, defecting turns out to be the more auspicious option. Therefore, no matter how auspicious cooperating might seem *before* you make up your mind about what to do, defecting is sure to look more auspicious *afterwards*—and this will be true no matter what decision you have made. Jeffrey proposes to uses this basic asymmetry—between what one decides and what one does—to argue that defection in PDT is the only rational course of action for an auspiciousness maximizer.

The case has already been made for an agent who is already certain about what she will decide. But what about agents who have yet to made up their minds? Here things get dicey. If you have not yet decided what to do, then the probabilities you assign to Δ_y^C and Δ_y^D will be far from one. This puts the auspiciousness values of C_y and D_y near those of $(C_y \& \Delta_y^C)$ and $(D_y \& \Delta_y^D)$ respectively, and since $\mathcal{V}(C \& \Delta_y^C) > \mathcal{V}(D \& \Delta_y^D)$, evidential decision theory tells you to choose cooperation. However, as soon as you make this choice, you will assign Δ_y^D a credence close to one, and as we have seen, you will then favor defection. Thus, the pursuit of good news forces you to make choices that you are certain to rue from the moment you make them—clearly something to avoid.

Jeffrey hopes to circumvent this difficulty by denying that evidential decision theory requires one to maximize auspiciousness as one *currently* estimates it. If you are savvy, he argues, you will realize that any choice you make will change some of your beliefs, thereby altering your estimates of auspiciousness. Thus, given that you want to make decisions that leave you better off for having made them, you should aim to maximize auspiciousness not as you currently estimate it, but as you *will* estimate it once your decision is made. You ought to, "choose for the person you expect to be when you have chosen,"[14] by maximizing expected utility computed relative to the personal probabilities you will have *after* having come to a firm decision about what to do. This is only possible if your choices conform to the maxim

- *Evidential Ratifiability*. An agent cannot rationally choose to perform A unless A is *ratifiable*, in the sense that $\mathcal{V}(A \& \triangle^A) \geq \mathcal{V}(B \& \triangle^A)$ for every act B under consideration.

This principle advises you to ignore your current views about the evidentiary merits of cooperating versus defecting, and to focus on maximizing future auspiciousness by making choices that you will regard as propitious from the epistemic perspective you will have once you have made them. Since in the presence of Screening, defection is the only such choice, the maxim of Evidential Ratifiability seems to provide an appropriately "evidentialist" rationale for defecting in PDT.

Unfortunately, though, the Screening condition need not always obtain. There are versions of PDT in which the actual performance of an act provides better evidence for some desired state than does the mere decision to perform it. This would happen, for example, if you and twin are bumblers who tend to have a similar problems carrying out your decisions. The fact that you were able to carry out a decision would then be evidentially correlated with you twin's act, and this correlation would not be screened off by the decision itself. In such cases the evidential ratifiability principle sanctions cooperation.[15] Therefore, the Jeffrey/Eells

[14] Jeffrey (1983) p. 16.

[15] Jeffrey, in his original published treatment of ratificationism (Jeffrey 1983), 20, gives this counterexample and credits it to Bas van Fraassen. Shin (1991b) treats cases in which the respective players' "trembles" are independent of each other.

strategy does not always provide a satisfactory evidentialist rationale for defecting in PDT. We regard this failure as reinforcing our contention that any adequate account of rational choice must recognize that decision makers have beliefs about causal or counterfactual relationships, beliefs that cannot be cashed out in terms of ordinary subjective conditional probabilities—in terms of conditional credences in non-causal propositions.

Ratificationism and Causality in the Theory of Games

Despite its failure, the Jeffrey/Eells strategy leaves the theory of rational choice an important legacy in the form of the Maxim of Ratifiability. We will see in this section that it is possible to assimilate Jeffrey's basic insight into casual decision theory and that so understood, it codifies a type of reasoning commonly employed in game theory. Indeed, the idea that rational players always play their part in a Nash equilibrium is a special case of ratificationism.

The notion of a ratifiable act makes sense within any decision theory with the resources for defining the expected utility of one act given the news that another will be chosen or performed. In causal decision theory the definition would be this:

$$\mathcal{U}(B/A) = \sum_{i=1}^{n} \rho((B \,\square\!\!\rightarrow S_i)/A)u(B, S_i)$$

Jeffrey's insight then naturally turns into the maxim,

- *Causal Ratifiability*. An agent cannot rationally choose to perform act A unless $\mathcal{U}(A/A) \geq \mathcal{U}(B/A)$ for every act B under consideration.

This says that a person should never choose A unless once she does, her expection of A's efficacity in bringing about desirable results is at least as great as that for any alternative to A. Notice that one no longer needs a "trembling hand" requirement to define the utility of one act conditional on another, since the conditional credence $\rho(B \,\square\!\!\rightarrow S/A)$ is well-defined even when A and B are incompatible. This means that an agent who is certain that she will perform A can still coherently wonder about how things *would* have gone *had* she done B—even though she can no longer wonder about how things are set to go if she's *going* to do B. This ability to assign utilities to actions that definitely will not be performed can be used to used to great advantage in game theory, for instance, when considering subgames and subgame perfection.

To see how an act might fail to be causally ratifiable, imagine yourself playing Matching Pennies with an opponent who you think can predict your move and will make a best response to it (see table). Neither pure act then turns out to be ratifiable. Suppose that you are strongly inclined to play [HEADS], so that $\rho(H_y)$ is close to one. Since your conditional probability for Twin playing heads given that you do

is high, it also follows that your subjective probability for H_t will be high. Thus by recognizing that you plan to play [HEADS], you give yourself evidence for thinking that Twin will also play [HEADS]. Note, however, this does nothing to alter the fact that you still judge Twin's actions to be causally independent of your own; your subjective probabilities still obey

	HEADS	TAILS
HEADS	−1, 1	1, −1
TAILS	1, −1	−1, 1

Matching Pennies

$$\rho\left(\left(H_y\Box\rightarrow H_t\right)/H_y\right) \approx \rho\left(\left(T_u\Box\rightarrow H_t\right)/H_y\right).$$

Since $\rho(Ht) \approx 1$, your overall position is this: because you are fairly sure that you will play heads, you are fairly sure that Twin *will* play heads too; but you remain convinced that she *would* play heads even if (contrary to what you expect) you were to play tails. Under these conditions, the conditional expected utility associated with [TAILS] larger than that associated with [HEADS] on the supposition that [HEADS] is played. That is to say, [HEADS] is unratifiable. Similar reasoning shows that tails is also unratifiable. In fact, the only ratifiable act this game is the mixture ½ [HEADS] + ½ [TAILS].[16]

It is no coincidence that this mixture also turns out to be the game's unique Nash equilibrium; there is a deep connection between ratifiable acts and game-theoretic equilibria. Take any two-person game, such as Matching Pennies, for which the unique Nash equilibrium consists of mixed strategies (which the players in fact adopt). If a player has predicted the other's strategy correctly and with certainty, then by playing the strategy she does, she maximizes her expected utility. But this isn't the only strategy that, given her credences, would maximize her expected utility; any other probability mixture of the same pure strategies would do so too. The

[16]Piccione and Rubinstein (1997) present another kind of case in which considerations of ratifiability may be invoked: the case of the "absent-minded driver" who can never remember which of two intersections he is at. One solution concept they consider (but reject) is that of being "modified multi-selves consistent". In our terms, this amounts to treating oneself on other occasions as a twin, selecting a strategy that is ratifiable on the following assumption: that one's present strategy is fully predictive of one's strategy in any other situation that is subjectively just like it. This turns out to coincide with the "optimal" strategy, the strategy one would adopt if one could choose in advance how to handle all such situations.

strategy she adopts is unique, though, in this way: it is the only strategy that could be ratifiable, given the assumption that her opponent has predicted her strategy and is playing a best response to it. It should be clear that this argument extends straightforwardly to the *n*-person case. In any Nash equilibrium, all players perform causally ratifiable acts.

	C_1	C_2
R_1	1, 1	0, 4
R_2	4, 0	−15, −15

Chicken

In fact, the least restrictive of all equilibrium solution concepts—Aumann's correlated equilibrium—can be understood as a straightforward application of the maxim of ratifiability.[17] Aumann sought to generalize the Nash equilibrium concept by relaxing the assumption—implicit in most previous game-theoretic thinking—that players in normal-form games believe that everyone acts independently of everyone else. Take a game of Chicken. The table shown might give the utilities of drivers traveling along different roads into a blind intersection, who must decide whether to stop and lose time (R_1 and C_1) or to drive straight through and risk an accident (R_2 and C_2). Let the players assume that they will choose, respectively, strategies $p \cdot R_1 + (1-p) \cdot R_2$ and $q \cdot C_1 + (1-q) \cdot C_2$. What credences will they give to the various joint pure actions they may end up performing?

Game theorists have traditionally assumed that the players will treat the chance devices that implement their mixed strategies as evidentially independent, ascribing the credences in Table 23.1. Aumann imagines that they might ascribe the credences of Table 23.2: To see how the correlations in Table 23.2 might arise, imagine a "managed" game of Chicken in which, to minimize the chances of a disastrous outcome, things have been arranged so that an impartial arbitrator will throw a fair die and illuminate a traffic light at the intersection according to the "Arbitrator's Scheme" shown in the table. If each player intends to heed the signal, stopping on red and driving through on green, and believes his opponent intends to do so as well, then both will have the correlated priors of Table 23.2. Aumann showed how to define an equilibrium solution concept even when players' acts are correlated in this way. An *adapted* strategy is a rule *s* which dictates a unique pure act $s(A)$ for

[17] More precisely, correlated equilibrium is the weakest equilibrium solution concept which assumes that all players have common beliefs. When this assumption is relaxed one obtains a subjectively correlated equilibrium. For details see Aumann (1974, 1987).

Table 23.1 Chicken with Independence

	C_1	C_2
R_1	pq	$p(1-q)$
R_2	$(1-p)q$	$(1-p)\ (1-q)$

Table 23.2 Chicken with Correlation

	C_1	C_2
R_1	$p\ +q-1$	$1-q$
R_2	$1-p$	0

each act A signalled by the arbitrator: for instance, "Stop on red and go on green" or "Run all lights." A correlated equilibrium is simply a pair of adapted strategies r and c such that, for all alternatives r^* and c^*, the following condition holds (call it *CE*):

	one	two	three	four	five	six
ROW	Red	Red	Red	Red	Green	Green
COL	Red	Red	Green	Green	Red	Red

Arbitrator's Scheme

$$\sum_{ij} \rho(R_i \& C_j) \cdot \mathcal{U}_{\text{row}}(r(R_i), C_j) \geq \sum_{ij} \rho(R_i \& C_j) \cdot \mathcal{U}_{\text{row}}(r^*(R_i), C_j)$$

$$\sum_{ij} \rho(R_i \& C_j) \cdot \mathcal{U}_{\text{col}}((R_i), c(C_j)) \geq \sum_{ij} \rho(R_i \& C_j) \cdot \mathcal{U}_{\text{col}}(R_i c^*(C_j))$$

Table 23.3 Equilibrium for Chicken with Independence

	C_1	C_2
R_1	⅓	⅓
R_2	⅓	0

Thus, r and c constitute a correlated equilibrium iff no player has reason to deviate from her adapted strategy given her beliefs and the signal she receives. The reader may verify that, when things are as described in Table 23.3, both players' resolving to obey the light is a correlated equilibrium, whereas there is no correlated equilibrium in which either player decides to run red lights.

Definition CE makes it appear as if the existence of correlated equilibria depends on the availability of external signaling mechanisms. This was the view presented by Aumann in (1974), but in (1987), he shows that acts themselves can serve as appropriate signals. Specifically, Aumann established that a common probability distribution[18] ρ defined over strategy combinations comprises a correlated equilibrium iff given any pair of acts R and C assigned positive probabilities, one has *CE*:*

$$\sum_i \rho(C_j/R) \cdot \mathcal{U}_{\text{row}}(R, C_j) \geq \sum_i \rho(C_j/R) \cdot \mathcal{U}_{\text{row}}(R^*, C_j)$$

$$\sum_i \rho(R_j/C) \cdot \mathcal{U}_{\text{col}}(R_i, C) \geq \sum_i \rho(R_j/C) \cdot \mathcal{U}_{\text{col}}(R_i, C^*)$$

for all alternatives R^* and C^*. CE* requires, then, that agents assign zero prior probability to any act, either of their own or of their adversaries', that does not maximize expected utility on the condition that it will be performed. Aumann regards this condition as "an expression of Bayesian rationality," since players satisfy it by maximizing expected utility at the time they act.

As a number of authors have noted, CE* is an application of the maxim of ratifiabilty.[19] It requires players to give zero credence to non-ratifiable acts. Hence for a group of players to end up in a correlated equilibrium, it is necessary and sufficient that all choose ratifiabile acts and expect others to do so as well. Aumann's "Bayesian rationality" thus coincides with the notion of rationality found in Jeffrey's ratificationism. More specifically, we contend that Aumann is requiring rational players to choose *causally* ratifiable actions.

[18]Note that such a distribution determines a unique mixed act for each player. Thus, it makes no difference whether one talks about the players' acts or the players' beliefs being in equilibrium.

[19]Shin (1991b), Skyrms (1990).

While Aumann has rather little to say about the matter, he clearly does not mean the statistical correlations among acts in correlated equilibrium to reflect causal connections. This is particularly obvious when external signaling devices are involved, for in such cases each player believes that his opponents would heed the arbitor's signal whether or not he himself were to heed it. Each player uses his knowledge of the arbitor's signal to him to make inferences about the signals given to others, to form beliefs about what his opponents expect him to do, and ultimately to justify his own policy of following the arbitor's signal. What he does not do is suppose that the correlations so discovered *would* continue to hold no matter what he decided to do. For example, if [ROW]'s credences are given in Table 23.3, it would be a mistake for him to run a red light in hopes of making it certain that [COLUMN] will stop; [COLUMN's action, after all, is determined by the signal she receives, not by what [ROW] does. Notice, however, that a straightforward application of evidential decision theory recommends running red lights in this circumstance—further proof that the view is untenable. In cases, then, where correlations are generated via external signaling, a causal interpretation of CE* clearly is called for. To make this explicit, we can rewrite CE* as the following condition *CE***:

$$\sum_j \rho((R \,\square\!\!\rightarrow C_j)/R) \cdot \mathcal{U}_{\text{row}}(R, C_j) \geq \sum_j \rho((R^* \,\square\!\!\rightarrow C_j)/R) \cdot \mathcal{U}_{\text{row}}(R^*, C_j)$$

$$\sum_j \rho((C \,\square\!\!\rightarrow R_i)/C) \cdot \mathcal{U}_{\text{col}}(R_i, C) \geq \sum_i \rho((C^* \,\square\!\!\rightarrow R_i)/C) \cdot \mathcal{U}_{\text{col}}(R_i, C^*)$$

This reduces to CE* on the assumption that [ROW] and [COLUMN] cannot influence each other's acts, so that both $\rho\left(\left(R * \;\square\!\!\rightarrow C_j\right)/R\right) = \rho\left(C_j/R\right)$ and $\rho\left(\left(C * \;\square\!\!\rightarrow R_i\right)/C\right) = \rho\left(R_i/C\right)$ hold for all R^* and C^*.

The situation is not appreciably different in cases where the correlation arises without any signaling device. Imagine playing [ROW] in the coordination game in Table 23.4, and consider the correlated equilibrium described by Table 23.5: These correlations need not have arisen through signaling. You might find the (R_1, C_1) equilibrium salient because it offers the vastly higher payoff, and, knowing that

Table 23.4 Correlation without Signaling

	C_1	C_2
R_1	25, 1	0, 0
R_2	0, 0	1, 2

Table 23.5 Equilibrium for Correlation without Signaling

	C_1	C_2
R_1	0.7	0.01
R_2	0.01	0.28

your opponent can appreciate its salience for you, you might conclude that she expects you to play it. That makes C_1 the better play for her, which reinforces your intention to play R_1 and your belief that she will play C_1, and so on. It would not be unreasonable under these conditions for the two of you to end up in the correlated equilibrium of Table 23.5. Still, there is no suggestion here that your initial inclination to play R_1 is somehow responsible causally for your opponent's credences. Her credences are what they are because she suspects that you are inclined to play R_1, but neither your decision to play R_1 nor your actually playing of R_1 is the cause of these suspicions – she develops them solely on the basis of her knowledge of the game's structure. As in the signaling case, the acts in a correlated equilibrium are evidentially correlated but causally independent.

This explicitly causal reading of Aumann's discussion helps to clear up a perplexing feature of correlated equilibria. CE only makes sense as a rationality constraint if agents are able to treat their own actions as bits of information about the world, for it is only then that the expressions appearing to the right of the "$\geq$" signs can be understood as giving utilities for the starred acts. As Aumann notes, his model (like that of Jeffrey before him)

> does away with the dichotomy usually perceived between uncertainty about acts of nature and of personal players . . . . In traditional Bayesian decision theory, each decision maker is permitted to make whatever decision he wishes, after getting whatever information he gets. In our model this appears not to be the case, since the decision taken by each decision maker is part of the description of the state of the world. This sounds like a restriction on the decision maker's freedom of action. (Aumann 1987), 8.

The problem here is that the utility comparisons in CE seem to portray acts as things that happen to agents rather than things they do. Moreover, it is not clear why an agent who learns that he is surely going to perform R should need to compare it with other acts that he is sure he will not perform. Aumann tries to smooth things over by suggesting that CE describes the perspective of agents, not as choosers, but as "outside observers." He writes,

> The "outside observer" perspective is common to all differential information models in economics . . . . In such models, each player gets some information or "signal"; he hears only the signal sent to him, not that of others. In analyzing his situation, [the] player must

> first look at the whole picture as if he were an outside observer; he cannot ignore the possibility of his having gotten a signal other than he actually got, even though he knows that he actually did not get such a signal. This is because the other players do not know what signal he got. [He] must take the ignorance of the other players into account when deciding on his own course of action, and he cannot do this if he does not explicitly include in the model signals other than the one he actually got. (Aumann 1987), 8.

The problem with this response is that it does not tell us how a player is supposed to use the knowledge gained from looking at things "externally" (i.e., as if he were not choosing) to help him with his "internal" decision (where he must choose how to act). The point is significant, since what's missing from the outside observer standpoint is the very thing that makes something a decision problem—the fact that a choice has to be made.

In our view, talk about outside observers here is misleading. Aumann's third-person, "outside observer" perspective is the first-person *subjunctive* perspective: a view on what *would* happen if *I were to do* something different. It is an advantage of causal decision theory that it allows one to assess the rationality of acts that certainly will not be performed in the same way as one assesses the rationality of any other option. One simply appeals to whatever facts about objective chances, laws of nature, or causal relations are required to imagine the "nearest" possible world in which the act is performed, the one most similar, in relevant respects, to the actual world in which one won't perform the act. One then sees what would ensue under those circumstances. Even, for instance, if one assigns zero credence to the proposition that one will intentionally leap off the bridge one is crossing, it still makes sense on the causal theory to speak of the utility of jumping. Indeed the abysmally low utility of this action is the principal reason why one is certain not perform it. When we interpret correlated equilibria causally, then, along the lines of CE**, there is no need for an "external" perspective in decision making. All decisions are made from the first-person subjunctive perspective of "What would happen if I performed that act?"

A number of other aspects of game-theoretic reasoning can likewise be analyzed in terms of causal ratifiability.[20] Some of the most interesting work in this area is due to Harper, who has investigated the ways in which causal decision theory and ratifiability can be used to understand extensive-form games (Harper 1986, 1991). This is a natural place for issues about causation to arise, since players' choices at early stages in extensive-form games can affect other players' beliefs—and thus their acts—at later stages.

Harper proposes using the maxim of ratifiability as the first step in an analysis of game-theoretic reasoning that is "eductive" or "procedural", an analysis that seeks to supply a list of rules and deliberative procedures that agents in states of indecision can use in order to arrive at an intuitively correct equilibrium choice.[21] Rational

[20] See Shin (1991b), for instance, for an interesting ratificationist gloss on Selten's notion of a "perfect" equilibrium.

[21] On the need for such "eductive" procedures see Binmore (1987, 1988).

players, he suggests, should choose actions that maximize their unconditional causal expected utility *from among the ratifiable alternatives*. (Harper regards cases where no ratifiable strategies exist as genuinely pathological.) The idea is not simply to choose acts that maximize unconditional expected utility, since these need not be ratifiable. Nor is it to choose acts with maximal expected utility on the condition that they are performed, since these may have low unconditional utility. Rather, one first eliminates unratifiable options, and then maximizes unconditional expected utility with what is left. In carrying out the second step of this process, each player imagines her adversaries choosing among their ratifiable options by assigning zero probability to any option that wouldn't be ratifiable.

Harper shows that both in normal and in extensive-form games, players who follow these prescriptions end up choosing the intuitively "right" act in a wide range of cases. In extensive-form games, his method produces choices that are in sequential equilibrium—and perhaps most interesting, the method seems to promote a strong form of "forward induction." To illustrate the latter point, and to get a sense of how Harper's procedure works in practice, consider the game of Harsanyi and Selten (1988) shown here. [ROW]'s act *A* yields fixed payoffs, whereas [ROW]'s act *B* leads to a strategic subgame with [COLUMN]. Harsanyi and Selten argue that (*C, e*) is the unique rational solution to the subgame, and they use backwards induction to argue that (*AC, e*) is the only rational solution in the full game. Their thought is that at the initial choice point [ROW] will know that (*C, e*) would be played if the subgame were reached, so he actually faces the "truncated game" as shown below, which makes *A* the only rational choice.

Kohlberg and Mertens (1986) have objected to this reasoning on the grounds that [ROW] can perform *B* as a way of signaling [COLUMN] that he has chosen *BD* (since it would be crazy to pass up *A* for *C*), and can thus force [COLUMN] to play *f* rather than *e*. Harsanyi and Selten respond by claiming that [COLUMN] would have to regard [ROW]'s playing *B* as a mistake since, "before deciding whether [ROW] can effectively signal his strategic intentions, we must first decide what strategies are rational for the two players in the subgame, and accordingly what strategy is the rational strategy for [ROW] in the truncation game" (Harsanyi and Selten 1988), 353 (see table). Thus, we have a dispute over what sorts of effects the playing of *B* would have on [COLUMN]'s beliefs at the second choice point, and thereby on her decision. This is just the sort of case where causal decision theory can be helpful.

Harper's procedure endorses Kohlberg's and Mertens' contention. Harsanyi and Selten's preferred act for [ROW] will be unratifiable, so long as each player knows that the other chooses only ratifiable options. For *A* to be ratifiable, it would at least have to be the case that

$$U(A/A) = 4 \geq \mathcal{U}(BD/A) = \rho((BD \,\square\!\!\rightarrow e)/A) \cdot 0 + \rho((BD \,\square\!\!\rightarrow f)/A) \cdot 10. \tag{23.14}$$

However, since [COLUMN] must choose at the third choice point knowing that [ROW] has played *B*, but without knowing whether he has played *C* or *D*, it follows that [ROW]'s credence for $BC \,\square\!\!\rightarrow f$ must be $\rho(f/B)$. Now at the final choice

point, [COLUMN] would know that [ROW] had chosen either *BC* or *BD* or some mixture of the two, and she would have to assume that the option chosen was ratifiable (if such an option is available). *BC* clearly cannot be ratifiable, since it is dominated by *A*. Harper also shows, using a somewhat involved argument, that no mixture of *BC* and *BD* can be ratifiable.[22] *BD*, however, is ratifiable, provided that $\rho(BD \;\square\!\!\rightarrow f/BD) = \rho(f/B) \geq {}^{4}/_{10}$. Thus, since only one of [ROW]'s B-acts can be ratifiable, [COLUMN] would have to assign it a probability of one if she were to find herself at the second choice point. [ROW], knowing all this and knowing that *f* is [COLUMN]'s only ratifiable response to *BD*, will indeed assign a high value to $\rho(f/B)$, viz. $\rho(f/B) = 1$. This in turn ensures that $\mathcal{U}(BD/A) > \mathcal{U}(A/A)$, making *A* unratifiable. *BD* is thus the only ratifiable solution to the Harsanyi/Selten game. Hence, if Harper is correct, it seems that Kohlberg and Mertens were right to reject the backwards induction argument and to think that [ROW] can use *B* to warn [COLUMN] of his intention to play *D*.

We hope this example gives the reader something of the flavor of Harper's approach. Clearly his proposal needs to be elaborated more fully before we will be able to make an informed judgment on its merits. We are confident, however, that any adequate understanding of game-theoretic reasoning will rely heavily on causal decision theory and the maxim of ratifiability.

Foundational Questions

Before any theory of expected utility can be taken seriously, it must be supplemented with a representation theorem that shows precisely how its requirements are reflected in rational preference. To prove such a theorem one isolates a small set of axiomatic constraints on preference, argues that they are requirements of rationality, and shows that anyone who satisfies them will automatically act in accordance with the theory's principle of expected utility maximization. The best known result of this type is found in Savage (1972), where it was shown that any agent whose preferences conform to the well-known Savage axioms will maximize expected utility, as defined by Eq. (23.1), relative to a (unique) probability ρ and a (unique up to positive linear transformation) utility *u*. Unfortunately, this result does not provide a fully satisfactory foundation for either CDT or ETD, because, as we have seen, Savage's notion of expected utility is ambiguous between a causal and an evidential interpretation. This leaves some important unfinished business for evidential and causal decision theorists, since each camp has an obligation to present a representation theorem that unambiguously captures its version of utility theory.

Evidential decision theorists were first to respond to the challenge. The key mathematical result was proved in Bolker (1966), and applied to decision theory in Jeffrey (1983). The Jeffrey/Bolker approach differs from Savage's in two significant

[22] Harper (1991), 293.

ways: First, preferences are defined over a σ-algebra of propositions that describe not only states of the world, but consequences and actions as well. Second, Savage's "Sure-thing" Principle (his postulate P3) is replaced by the weaker

- *Impartiality Axiom*: Let X, Y and Z be non-null propositions such that (a) Z is incompatible with both X and Y, (b) X and Y are indifferent in the agent's preference ordering, and (c) $(X \vee Z)$ is indifferent with $(Y \vee Z)$ but not with Z. Then, $(X \vee Z^*)$ must be indifferent with $(Y \vee Z^*)$ for any proposition Z^* incompatible with both X and Y.

In the presence of the other Bolker/Jeffrey axioms, which do not differ substantially from those used by Savage, Impartiality guarantees that the agent's preferences can be represented by a function that satisfies Eq. (23.3).[23] It also ensures that the representation will be *partition independent* in the sense that, for any partitions $\{X_i\}$ and $\{Y_j\}$ and any act A, one will always have $\mathcal{V}(A) = \sum \rho(X_i/A)\mathcal{V}(A\&X_i) = \sum \rho(Y_j/A)V(A\&Y_j)$. Thus, in EDT it does not matter how one chooses the state partition relative to which expected utilities are computed.

This contrasts sharply with Savage's theory. Formula (23.1) shows how to compute expected utilities with respect to a single partition of states, but it gives no guarantee that different choices of partitions yield the same value for $\mathcal{V}(A)$. As a consequence, Savage had to place restrictions on the state partitions that could be legitimately employed in well-posed decision problems. Strictly speaking, he said, his axioms only apply to *grand-world* decisions whose act and state partitions slice things finely enough to ensure that each act/state pair produces a consequence that is sufficiently detailed to decide every question the agent cares about. Thus, on the official view, we can only be confident that (23.1) will yield the correct value for $\mathcal{V}(A)$ when applied to state partitions in "grand-world" decisions. Savage recognized this as a significant restriction on his theory, since owing to the extreme complexity of grand-world decisions, no actual human being can ever contemplate making one. He tried to make this restriction palatable by suggesting that his axioms might be usefully applied to certain "small-world" decisions, and expressing the hope that there would be only a "remote possibility" of obtaining values for $\mathcal{V}(A)$ inconsistent with those obtained in the grand-world case. This hope, however, was never backed-up by any rigorous proof. We take it as a mark in favor of EDT that it is can solve this "problem of the small-world" by giving such a proof.

The easiest way to prove a representation result for CDT is to co-opt Savage's theorem by stipulating that well-posed decision problems must be based on partitions of states that are certain to be causally independent of acts, and then imposing Savage's axioms on such problems. A number of causal decision

[23] This representation will not be unique (except in the rare case where $\mathcal{V}$ is unbounded), for, as a simple calculation shows, if the function $\mathcal{V}(A) = \sum \rho(S_i/A)u(A, S_i)$ represents a preference ordering, and if k is such that $1 + k\mathcal{V}(X) > 0$ for all propositions X in the algebra over which $\mathcal{V}_k(A) = \sum \rho k(S_i/A)u_k(A, S_i)$ is defined, then $V_k(A) = \sum \rho k\,(S_i/A)\,u_k\,(A, S_i)$ will also represent the ordering, when $pk(X) = \rho(X)(1 + k\mathcal{V}(X))$ and $\mathcal{V}_k(X) = [\mathcal{V}(X)(1 + k)]/(1 + k\mathcal{V}(X))$.

theorists have endorsed the view that there is a "right" partition of states to be used for computing instrumental expected utility.[24] The Lewis formulation of CDT in terms of dependency hypotheses mentioned in the section "Conditionals and their probabilities" above is an example of this strategy (Lewis 1981). Minor intramural squabbles aside, the idea is that each element in the privileged partition, often denoted $\mathbf{K} = \{K_j\}$ following Skyrms (1980), should provide a maximally specific description of one of the ways in which things that the agent cares about might depend on what she does.[25] It is characteristic of such partitions that they will be related to the agent's subjective probability by[26]

- *Definiteness of Outcome*: For any proposition O that the agent cares about (in the sense of not being indifferent between O and $\neg O$), any action A, and any $K_j \in \mathbf{K}$, either $\rho((A \,\Box\!\!\rightarrow O)/K_j) = 1$ or $\rho((A \,\Box\!\!\rightarrow \neg O)/K_j) = 1$.
- *Instrumental Act Independence*: $\rho([A \,\Box\!\!\rightarrow K_j] \leftrightarrow K_j) = 1$ for all acts A and states K_j.

The first of these ensures that $u\,(A,\, S_i)$ has the same value for each Savage-state S_i in K_j, and thus that

$$\mathcal{U}(A) = \sum_i \rho(A \,\Box\!\!\rightarrow S_i) u(A, S_i) = \sum_j \rho(A \,\Box\!\!\rightarrow K_j) \mathcal{U}(A, K_j)$$

The second condition then guarantees that $\mathcal{U}(A) = \sum_j \rho(K_j)\mathcal{U}(A, K_j)$. Since this equation has the form (23.5), it follows that if there exists a partition $\mathbf{K}$ that meets these two requirements, then one can appropriately apply the Savage axioms to actions whose outcomes are specified in terms of it. Fortunately, the required partition is certain to exist, because it is always possible to find a $\mathbf{K} = \{K_i\}$ such that (i) K_j entails either $(A \,\Box\!\!\rightarrow O)$ or $(A \,\Box\!\!\rightarrow \neg O)$ for every O the agent cares about, and (ii) $([A \,\Box\!\!\rightarrow K_j] \leftrightarrow K_j)$ is a truth of logic for all A and K_j.[27] The general existence of such partitions lets us use Savage's representation theorem as a foundation for CDT, subject to the proviso that the Savage axioms should only be applied to decisions framed in terms of a $\mathbf{K}$-partition.

The trouble with this strategy is that the partition dependence of Savage's theory is carried over into CDT. The need for a partition-independent formulation of CDT

[24] See, for example, Skyrms (1980), Lewis (1981), Armendt (1986).

[25] Notice that states are being viewed here as functions from acts to outcomes, whereas acts are taken as unanalyzed objects of choice (that is, as propositions the agent can make true or false as she pleases). This contrasts with Savage's well-known formalization in which acts are protrayed as functions from states to outcomes, and states are left as unanalyzed objects of belief. Less hangs on this distinction than one might think. When one adopts the perspective of Jeffrey (1967, 1983) and interprets both states and actions as propositions, and views outcomes as conjunctions of these propositions, the two analyses become interchangeable.

[26] Here we are following Gibbard (1986).

[27] An explicit construction of $\mathbf{K}$ can be found in Gibbard and Harper (1978).

has been argued by Sobel (1989), and Eells (1982) has suggested that EDT should be preferred to CDT on this basis alone. The main difficulty is the problem of "small-worlds", which threatens to make the theory inapplicable, in the strictest sense, to all the decision problems people actually consider. Gibbard (1986) goes a certain distance toward alleviating this difficulty by, in effect, showing how to find the smallest **K**-partition for a given decision problem, but his partition is still rather "grand", and a fully partition-invariant version of CDT would still be desirable.

Armendt (1986) proves a representation result that does provide a formulation of CDT that is partition-independent, even though it does not do away with the notion of a **K**-partition. He takes the conditional decision theory of Fishburn (1974) as his starting point. The basic concept here is that of the utility of a prospect X on the hypothesis that some condition C obtains. If $\{C_1, C_2, \ldots, C_n\}$ is a partition of C, these conditional utilities are governed by the (partition independent) equation:

$$\mathcal{U}(X/C) = \sum \rho(C_i/C)\mathcal{U}(X/C_i) \tag{23.15}$$

which shows how X's utility given C depends on its utilities given the various ways in which C might be true. Notice that (23.15) allows for a distinction between an act A'*s* unconditional utility and its utility conditional on its own performance. These are given respectively by

$$\mathcal{U}(A) = \mathcal{U}(A/A \vee \neg A) = \sum \rho(S_i)\mathcal{U}(A/S_i)$$

$$\mathcal{U}(A/A) = \sum \rho(S_i/A)\mathcal{U}(A/A \,\&\, S_i)$$

where the state partition may be chosen arbitrarily. In a suggestion that bears some similarities to Jeffrey's ratificationist proposal, Armendt argues that decision problems in which an agent's unconditional preference for A differs from her preference for A conditional on itself are just the kinds of cases in which A'*s* auspiciousness diverges from its instrumental expected utility. This suggests a way of characterizing **K**-partitions directly in terms of the agent's preferences. Armendt's thought is that the elements of an appropriate **K**-partition should "screen-off" differences in value between unconditional A and A-conditional-on-A, so that the agent is indifferent between A given K_i and A given $(A\&K_i)$ for every i. When this is so, we will have $\sum \rho(K_i)\mathcal{U}(A/K_i) = \sum \rho(K_i/A)\mathcal{U}(A/K_i)$, and Armendt shows that the conditional utilities $\mathcal{U}(A/K_i)$ can be eliminated in favor of the unconditional news values $\mathcal{V}(A\&K_i)$ as long as there exists at least one partition of "consequences" $\mathbf{O} = \{O_j\}$ such that the agent's unconditional utility for $(A\&O_j\&K_i)$ is equal to her utility for A conditional on $(A\&O_j\&K_i)$. When such a **K** and **O** exist, we have $\mathcal{U}(A) = \sum \rho(K_i)\mathcal{V}(A \,\&\, K_i)$, which is precisely the condition in which CDT and EDT coincide. What Armendt shows, then, is that an appropriate representation theorem for CDT can be obtained by supplementing Fishburn's conditional decision theory with the assumption that every act A can be associated with a **K**-partition such that $A/K_i \approx A/(A\&K_i)$ for all i, and a partition of consequences **O** (dependent on A and **K**) such that $(A\&O_j\&K_i) \approx A/(A\&O_j\&K_i)$.

This is a nice result. Since Fishburn's theory is partition independent, it follows that CDT will be as well, provided that at least one pair of partitions **K, O** exist for each act A. The crucial questions are whether such partitions do exist in general, and whether we should think that the condition that defines **K** really does guarantee that $\mathcal{U}(A)$ and $\mathcal{V}(A)$ coincide. On this latter point we have our doubts, but even if it is granted, it seems unlikely to us, in the absence of further argument, that the appropriate **K**-partitions will exist in all cases where we would want to apply CDT. Indeed, it would be useful to have a representation theorem for CDT that does not need to assume the existence of any special partition of states or any canonical form for a decision problem to take.

A representation theorem with these desirable features is proven in Joyce (1999). Joyce sees both EDT and CDT as instances of an abstract conditional expected utility whose basic concept is that of the utility of an act A on the supposition that some condition C obtains. Joyce begins by characterizing supposition, or provisional belief revision, in terms sufficiently general to subsume Bayesian conditioning and Lewis's imaging as special cases. Given a subjective probability ρ defined over a σ-algebra Ω, and a distinguished subset $\mathcal{C}$ of *conditions* in Ω,[28] a *supposition for* ρ *relative to* C is a function $\rho(\bullet \mid \bullet)$ from $\Omega \times \mathcal{C}$ into the real numbers that satisfies

(a) $\rho(\bullet \mid C)$ is a countably additive probability on Ω for every $C \in \mathcal{C}$.
(b) $\rho(C \mid C) = 1$ for all $C \in \mathcal{C}$.
(c) $\rho(X \mid C \vee \neg C) = \rho(X)$ for all $X \in \Omega$.
(d) $\rho(X \mid B\&C) \geq \rho(X\&B \mid C)$ for all $X \in \Omega$ whenever $(B\&C) \in \mathcal{C}$.
(e) Let B and C be mutually incompatible conditions in $\mathcal{C}$. Then if one has $\rho(X \mid B) \geq \rho(X \mid C)$, then one has $\rho(X \mid B \vee C) \geq \rho(X \mid C)$, with equality if either $\rho(X \mid B) = \rho(X \mid C)$ or $\rho(B \mid B \vee C) = 0$.

The reader can verify that the ordinary conditional probability $\rho(X/C)$—that is, $\rho(X\&C)/\rho(C)$—is a supposition for p relative to $\mathcal{C} = \{C \in \Omega : \rho(C) > 0\}$. In fact, for any set of conditions C (even those containing conditions with zero prior probability), one can show that any map $\rho(\bullet \mid \bullet)$ defined on $\Omega \times \mathcal{C}$ that satisfies (a) – (c) plus

- *Bayes's Law*: $\rho(B \mid C)\rho(X \mid B\&C) = \rho(X \mid C)\rho(B \mid X\&C)$ whenever we have $(B\&C), (X\&C) \in C$.

is a supposition.[29] The imaging function $p^C = \rho(C \,\square\!\!\rightarrow\, \bullet)$ associated with a similarity relation among possible worlds is also a supposition, but not typically

[28] The set C always takes the form $C = \Omega \sim I$, where the *ideal* I is a collection of Ω-propositions that contains the contradictory event $(X\&\neg X)$, is closed under countable disjunctions, and which contains $(X\&Y)$ whenever $X \in I$ and $Y \in \Omega$.

[29] These Bayesian suppositions were defined in Renyi (1955), and have come to be called "Popper measures" in the philosophical literature, after Popper (1934). Interested readers may consult van Fraassen (1995), Hammond (1994), and McGee (1994) for informative discussions of Popper measures.

one that satisfies Bayes's Law. There are also suppositions that are neither Bayesian nor instances of imaging.

Joyce, impressed by the need for partition-independent formulations of utility theories, uses the notion of a supposition to define an abstract conditional expected utility to be thought of as "utility under a supposition". Its (partition independent) basic equation is

$$\begin{aligned}\mathcal{V}(A|C) &= \sum_i \frac{\rho(S_i \,\&\, A|C)}{\rho(A|C)} u(A, S_i) \\ &= \sum_i \frac{\rho(X_i \,\&\, A|C)}{\rho(A|C)} \mathcal{V}(A \,\&\, X_i|C) \text{ for any partition } \{X_i\},\end{aligned} \tag{23.16}$$

where $\rho(\bullet \mid \bullet)$ might be any supposition function for ρ defined relative to a set of conditions C that contains propositions describing all an agent's actions (as well as other things). Since (23.16) is just EDT's (23.3) with $\rho(\bullet \mid C)$ substituted in for $\rho(\bullet)$, $\mathcal{V}(A|C)$ gives A's auspiciousness on the supposition that condition C obtains, where this supposition may be any provisional belief revision that satisfies (a) – (e).

As with Fishburn's theory, there is no guarantee that A's unconditional utility, which is now just $\mathcal{V}(A)$, will coincide with its utility conditional on itself,

$$\mathcal{V}(A|A) = \sum_i \rho(S_i|A) u(A, S_i)$$

The sole exception occurs when $\rho(\bullet \mid \bullet)$ is Bayesian, for in that case we have $\mathcal{V}(A) = \mathcal{V}(A \mid A)$, because $\rho(Si \mid A) = \rho(S_i \& A)/\rho(A)$ for all i. Given that $\mathcal{V}(A)$ and $\mathcal{V}(A \mid A)$ can differ in general, it becomes a live question whether a decision maker should choose acts that maximize her unconditional expected utility or choose acts that maximize expected utility conditional on the supposition that they are performed. Joyce argues, on grounds having nothing to do with the conflict between CDT and EDT, that a choiceworthy action is always one that maximizes expected utility on the condition that it is performed. The rational decision maker's objective, in other words, should always be to choose an A such that $\mathcal{V}(A \mid A) \geq \mathcal{V}(B \mid B)$ for all alternatives B. Neither evidential nor causal decision theorists will dispute this point, since the former endorse the prescription to maximize $\mathcal{V}(A \mid A)$ when the supposition function is $\rho(A \mid C) = \rho_c(A)$, which makes $\mathcal{V}(A|A)$ equal $A's$ auspiciousness, and the latter endorse it when $\rho(A \mid C) = \rho(C \,\square\!\!\rightarrow A)$, which makes $\mathcal{V}(A \mid A)$ equal A's instrumental expected utility. Thus, EDT and CDT are both instances of abstract conditional utility theory. The difference between them has to do not with the basic form of the utility function or with the connection between expected utility and choiceworthiness, but with the correct type of supposition to use in decision making contexts.

Once we recognize this, it becomes clear that the problem of proving a representation theorem for CDT can be subsumed under the more general problem of proving a representation theorem for an abstract conditional utility theory. And, since the

function $\mathcal{V}(\bullet \mid C)$ obeys Eq. (23.3) relative to any fixed condition C, this latter problem can be solved by extending the Bolker/Jeffrey axioms for unconditional preferences to conditional preferences, and showing that anyone who satisfies the axioms is sure to have conditional preferences that can be represented by some function $\mathcal{V}(A \mid C)$ of form (23.16) that is defined relative to a supposition function $\rho(\bullet \mid \bullet)$ for her subjective probability ρ.

Joyce was able to accomplish this. We refer to reader to (1999) for the details, which turn out to be rather complicated, but the basic idea is straightforward. One starts by imagining an agent with a system of conditional preferences of the form: X on the supposition that B is weakly preferable to Y on the supposition that C, written $X \mid B \geq Y \mid C$. One assumes that this ranking obeys the usual axioms: transitivity, connectedness, a continuity principle, an Archimedean axiom, and so on. One also requires each *section* of the ranking $X \mid C \geq Y \mid C$, for C fixed, to satisfy the Bolker/Jeffrey axioms. Bolker's theorem then ensures that each section will be associated with a family r_C of $(\mathcal{V}, \rho_C)$ pairs that satisfy Eq. (23.3) and that represent $X \mid C \geq Y \mid C$. Different r_C-pairs will be related by the equations $\rho_C{}^*(X) = \rho_C(X)[1 + k\mathcal{V}_C(X)]$ and $\mathcal{V}_C{}^*(X) = \mathcal{V}_C(X)[(k+1)/(1 + k\mathcal{V}_C(X)]$, where k is any real number such that $[1 + k\mathcal{V}(X)] > 0$ for all propositions X such that $\mathcal{V}(X)$ is defined. The trick to proving a representation theorem for conditional decision theory is to find further constraints on conditional preferences under which it is possible to select a unique $(\mathcal{V}, P_C)$ pair from each r_C in such a way that $\mathcal{V}(X) \geq \mathcal{V}(X)$ is guaranteed to hold whenever $X \mid B \geq Y \mid C$. The main axiom that is needed is the following generalization of Impartiality:

- Let X_1, X_2, and X_3, and Y_1, Y_2, and Y_3, be mutually incompatible, and suppose that

$$X_1 \,|B \approx Y_1|\, C > (X_1 \vee X_2)\,|B \approx (Y_1 \vee Y_2)|\, C > X_2\,|B \approx Y_2|\, C \tag{23.17}$$

  holds for some conditions B and C. Then, if

$$X_1|B \approx Y_1|C \not\approx X_3|B \approx Y_3|C \not\approx (X_1 \vee X_3)|B \approx (Y_1 \vee Y_3)|C \tag{23.18}$$

  then

$$(X_1 \vee X_2 \vee X_3)\,|B \approx (Y_1 \vee Y_2 \vee Y_3)|\, C. \tag{23.19}$$

This is not as complicated as it looks. If clause (23.17) holds, the only sort of conditional utility that will represent $X \mid C \geq Y \mid B$ will be one in which $\rho(X_1 \mid B)/\rho(X_1 \vee X_2 \mid B) = \rho(Y_1 \mid C)/\rho(Y_1 \vee Y_2 \mid C)$. Likewise, if (23.18) holds, then the representation must be one in which $\rho(X_1 \mid B)/\rho(X_1 \vee X_3 \mid B) = \rho(Y_1 \mid C)/\rho(Y_1 \vee Y_3 \mid C)$. Together these two equalities entail that $\rho(X_1 \mid B)/\rho(X_1 \vee X_2 \vee X_3 \mid B) = \rho(Y_1 \mid C)/\rho(Y_1 \vee Y_2 \vee Y_3 \mid C)$, and this is just what (23.19) guarantees.

Using this axiom as the main formal tool, Joyce is able to construct a full conditional expected utility representation for the ranking $X \mid C \geq Y \mid B$. By

adding further conditions, one can ensure either that the representation's supposition function will be Bayesian or that it will arise via imaging from some similarity relation among possible worlds. In this way, both EDT and CDT are seen to have a common foundation in the abstract theory of conditional expected utility.

Conclusion

While the classical theory of Ramsey, de Finetti and Savage remains our best account of rational choice, its development has yet to be completed. An adequate theory should explain the role in decision making of causal thinking. True, a decision theory that did without causal propositions would have been nice: Cause and effect have long puzzled and divided philosophers and scientists, and theoretical discussion of causal methodology remains underdeveloped.[30] In decision making, though, we are stuck with causal thinking. Rational choice always involves judgements of how likely an option is to have various desirable consequences—and such judgements, we have argued, require the decision maker to have views, explicit or implicit, about causal or counterfactual relationships. Nothing else can substitute for these causal beliefs. The conditional credences employed by evidential decision theory cannot, because they are unable to distinguish causation from correlation. More refined "screening" techniques, while better at capturing causal connections, fail to apply in a important class of cases. To specify the kind of case to which they do apply, we must, one way or another, invoke causal relations.

We should not find this need for causal notions distressing. We draw causal conclusions all the time, after all, and scientists are able to glean causal tendencies from experiment and statistical data, using methods of high sophistication. Still, no one theory of causal notions has the precision and orthodox status of, say, the standard theory of subjective probability. Thus, an adequate decision theory, if we are right, must depend on new advances in our understanding of causation and its relation to rational belief. We might have wished that theoretical life had turned out easier, but as matters stand, important work on the foundations of utility theory remains to be done.

References

Adams, E. W. (1975). *The logic of conditionals*. Dordrecht: Reidel.
Armendt, B. (1986). A foundation for causal decision theory. *Topoi, 5*, 3–19.
Aumann, R. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics, 1*, 67–96.

[30] Again, giant steps have been taken in this area since this article first appeared, especially by Spirtes et al. (1993) and Pearl (2000).

Aumann, R. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica, 55*, 1–18.
Binmore, K. (1987). Modeling rational players I. *Economics and Philosophy, 3*, 179–212 (Reprinted, Binmore 1990).
Binmore, K. (1988). Modeling rational players II. *Economics and Philosophy, 4*, 9–55 (Reprinted Binmore 1990).
Binmore, K. (1990). *Essays on the foundations of game theory*. Oxford: Blackwell.
Bolker, E. (1966). Functions resembling quotients of measures. *Transactions of the American Mathematical Society, 124*, 292–312.
Eells, E. (1982). *Rational decision and causality*. Cambridge: Cambridge University Press.
Fine, K. (1975). Review of Lewis [24]. *Mind, 84*, 451–458.
Fishburn, P. (1974). A mixture-set axiomatization of conditional subjective expected utility. *Econometrica, 41*, 1–25.
Gibbard, A. (1986). A characterization of decision matrices that yield instrumental expected utility. In L. Daboni, A. Montesano, & M. Lines (Eds.), *Recent developments in the foundations of utility and risk theory* (pp. 139–148). Dordrecht: Reidel.
Gibbard, A., & Harper, W. L. (1978). Counterfactuals and two kinds of expected utility. In C. A. Hooker, J. J. Leach, & E. F. McClennen (Eds.), *Foundations and applications of decision theory* (Vol. I). Dordrecht: Reidel.
Hammond, P. (1994). Elementary non-Archimedean representations of probability for decision theory and games. In P. Humphries (Ed.), *Patrick Suppes: Scientific philosopher* (Vol. 1, pp. 25–61). Dordrecht: Kluwer Academic Publishers.
Harper, W. (1986). Mixed strategies and ratifiability in causal decision theory. *Erkenntnis, 24*, 25–26.
Harper, W. (1991). Ratifiability and refinements in two-person noncooperative games. In M. Bacharach & S. Hurley (Eds.), *Foundations of decision theory* (pp. 263–293). Oxford: Basil Blackwell.
Harsanyi, J., & Selten, R. (1988). *A general theory of equilibrium selection in games*. Cambridge: MIT Press.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.
Horgan, T. (1981). Counterfactuals and Newcomb's problem. *Journal of Philosophy, 68*, 331–356.
Howard, J. V. (1988). Cooperation in the prisoner's dilemma. *Theory and Decision, 24*, 203–213.
Jeffrey, R. (1967). *The logic of decision*. New York: McGraw Hill.
Jeffrey, R. (1983). *The logic of decision* (2nd ed.). Chicago: University of Chicago Press.
Joyce. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.
Kohlberg, E., & Mertens, J. (1986). On the strategic stability of equilibria. *Econometrica, 54*, 1003–1037.
Lewis, D. K. (1973). *Counterfactuals*. Cambridge: Harvard University Press.
Lewis, D. K. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review, 85*, 297–315 (Reprinted, Lewis 1986, 133–152).
Lewis, D. K. (1979). Counterfactual dependence and time's arrow. *Noûs, 13*, 455–476 (Reprinted, Lewis 1986, 32–52).
Lewis, D. K. (1979). Prisoner's dilemma is a Newcomb problem. *Philosophy and Public Affairs, 8*, 235–240 (Reprinted, Lewis 1986, 299–304).
Lewis, D. K. (1980). A subjectivist's guide to objective chance. In Richard C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. 2). Berkeley: University of California Press. Reprinted (Lewis 1986), 83–113.
Lewis, D. K. (1981). Causal decision theory. *Australasian Journal of Philosophy, 59*, 5–30 (Reprinted, Lewis 1986, 305–337).
Lewis, D. K. (1986). *Philosophical papers* (Vol. 2). Oxford: Oxford University Press.
Luce, R. D., & Krantz, D. H. (1971). Conditional expected utility. *Econometrica, 39*, 253–271.
McGee, V. (1994). Learning the impossible. In E. Eells & B. Skyrms (Eds.), *Probability and conditionals* (pp. 179–197). Cambridge: Cambridge University Press.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel*. Dordrecht-Holland: Reidel.

Piccione, M., & Rubinstein, A. (1997). On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior, 20*, 3–24.

Popper, K. (1934). *Logik der Forschung.* Vienna: Springer. Translated as *The logic of scientific discovery* (London: Hutchinson, 1959).

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Renyi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarium Hungaricae, 6*, 285–335.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling, 7*, 1393–1512.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688–701.

Savage, L. J. (1972). *The foundations of statistics.* New York: Dover (First edition 1954).

Shin, H. S. (1991a). A reconstruction of Jeffrey's notion of ratifiability in terms of counterfactual beliefs. *Theory and Decision, 31*, 21–47.

Shin, H. S. (1991b). Two notions of ratifiability and equilibrium in games. In M. Bacharach & S. Hurley (Eds.), *Foundations of decision theory* (pp. 242–262). Oxford: Basil Blackwell.

Simon, H. A. (1957). *Models of man*. New York: Wiley.

Skyrms, B. (1980). *Causal necessity*. New Haven: Yale University Press.

Skyrms, B. (1984). *Pragmatics and empiricism*. New Haven: Yale University Press.

Skyrms, B. (1990). Ratifiability and the logic of decision. *Midwest Studies in Philosophy, 15*, 44–56.

Sobel, J. H. (1989). Partition theorems for causal decision theories. *Philosophy of Science, 56*, 70–95.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search* (Lecture Notes in Statistics, Vol. 81). New York: Springer. ISBN: 978-1-4612-7650-0 (Print) 978-1-4612-2748-9 (Online).

Stalnaker, R. (1968). A theory of conditionals. In *Studies in logical theory* (American philosophical quarterly monograph series 2). Oxford: Blackwell.

Stalnaker, R. (1972). Letter to David Lewis. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance, and time* (pp. 151–152). Dordrecht: Reidel.

Stalnaker, R. (1981). Letter to David Lewis of May 21, 1972. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance, and time*. Dordrecht-Holland: Reidel.

Stalnaker, R., & Thomason, R. (1970). A semantic analysis of conditional logic. *Theoria, 36*, 23–42.

Van Fraassen, B. (1995). Fine-grained opinion, probability, and the logic of belief. *Journal of Philosophical Logic, 24*, 349–377.

Chapter 24
Advances in Prospect Theory: Cumulative Representation of Uncertainty

Amos Tversky and Daniel Kahneman

Expected utility theory reigned for several decades as the dominant normative and descriptive model of decision making under uncertainty, but it has come under serious question in recent years. There is now general agreement that the theory does not provide an adequate description of individual choice: a substantial body of evidence shows that decision makers systematically violate its basic tenets. Many alternative models have been proposed in response to this empirical challenge (for reviews, see Camerer 1989; Fishburn 1988; Machina 1987). Some time ago we presented a model of choice, called prospect theory, which explained the major violations of expected utility theory in choices between risky prospects with a small number of outcomes (Kahneman and Tversky 1979; Tversky and Kahneman 1986). The key elements of this theory are (1) a value function that is concave for gains, convex for losses, and steeper for losses than for gains, and (2) a nonlinear transformation of the probability scale, which overweights small probabilities and underweights moderate and high probabilities. In an important later development, several authors (Quiggin 1982; Schmeidler 1989; Yaari 1987; Weymark 1981) have advanced a new representation, called the rank-dependent or the cumulative functional, that transforms cumulative rather than individual probabilities. This article presents a new version of prospect theory that incorporates the cumulative functional and extends the theory to uncertain as well to risky prospects with any number of outcomes. The resulting model, called cumulative prospect theory,

Amos Tversky was deceased at the time of publication.

A. Tversky (deceased)
Stanford University, Stanford, CA, USA

D. Kahneman (✉)
Princeton University, Princeton, NJ, USA
e-mail: kahneman@princeton.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_24

combines some of the attractive features of both developments (see also Luce and Fishburn 1991). It gives rise to different evaluations of gains and losses, which are not distinguished in the standard cumulative model, and it provides a unified treatment of both risk and uncertainty.

To set the stage for the present development, we first list five major phenomena of choice, which violate the standard model and set a minimal challenge that must be met by any adequate descriptive theory of choice. All these findings have been confirmed in a number of experiments, with both real and hypothetical payoffs.

Framing Effects The rational theory of choice assumes description invariance: equivalent formulations of a choice problem should give rise to the same preference order (Arrow 1982). Contrary to this assumption, there is much evidence that variations in the framing of options (e.g., in terms of gains or losses) yield systematically different preferences (Tversky and Kahneman 1986).

Nonlinear Preferences According to the expectation principle, the utility of a risky prospect is linear in outcome probabilities. Allais's (1953) famous example challenged this principle by showing that the difference between probabilities of .99 and 1.00 has more impact on preferences than the difference between 0.10 and 0.11. More recent studies observed nonlinear preferences in choices that do not involve sure things (Camerer and Ho 1991).

Source Dependence People's willingness to bet on an uncertain event depends not only on the degree of uncertainty but also on its source. Ellsberg (1961) observed that people prefer to bet on an urn containing equal numbers of red and green balls, rather than on an urn that contains red and green balls in unknown proportions. More recent evidence indicates that people often prefer a bet on an event in their area of competence over a bet on a matched chance event, although the former probability is vague and the latter is clear (Heath and Tversky 1991).

Risk Seeking Risk aversion is generally assumed in economic analyses of decision under uncertainty. However, risk-seeking choices are consistently observed in two classes of decision problems. First, people often prefer a small probability of winning a large prize over the expected value of that prospect. Second, risk seeking is prevalent when people must choose between a sure loss and a substantial probability of a larger loss.

Loss Aversion One of the basic phenomena of choice under both risk and uncertainty is that losses loom larger than gains (Kahneman and Tversky 1984; Tversky and Kahneman 1991). The observed asymmetry between gains and losses is far too extreme to be explained by income effects or by decreasing risk aversion.

The present development explains loss aversion, risk seeking, and nonlinear preferences in terms of the value and the weighting functions. It incorporates a framing process, and it can accommodate source preferences. Additional phenomena that lie beyond the scope of the theory—and of its alternatives—are discussed later.

The present article is organized as follows. Section "Cumulative prospect theory" introduces the (two-part) cumulative functional; section "Relation to previous work" discusses relations to previous work; and section "Values and weights"

describes the qualitative properties of the value and the weighting functions. These properties are tested in an extensive study of individual choice, described in section "Experiment", which also addresses the question of monetary incentives. Implications and limitations of the theory are discussed in section "Discussion". An axiomatic analysis of cumulative prospect theory is presented in the appendix.

Theory

Prospect theory distinguishes two phases in the choice process: framing and valuation. In the framing phase, the decision maker constructs a representation of the acts, contingencies, and outcomes that are relevant to the decision. In the valuation phase, the decision maker assesses the value of each prospect and chooses accordingly. Although no formal theory of framing is available, we have learned a fair amount about the rules that govern the representation of acts, outcomes, and contingencies (Tversky and Kahneman 1986). The valuation process discussed in subsequent sections is applied to framed prospects.

Cumulative Prospect Theory

In the classical theory, the utility of an uncertain prospect is the sum of the utilities of the outcomes, each weighted by its probability. The empirical evidence reviewed above suggests two major modifications of this theory: (1) the carriers of value are gains and losses, not final assets; and (2) the value of each outcome is multiplied by a decision weight, not by an additive probability. The weighting scheme used in the original version of prospect theory and in other models is a monotonie transformation of outcome probabilities. This scheme encounters two problems. First, it does not always satisfy stochastic dominance, an assumption that many theorists are reluctant to give up. Second, it is not readily extended to prospects with a large number of outcomes. These problems can be handled by assuming that transparently dominated prospects are eliminated in the editing phase, and by normalizing the weights so that they add to unity. Alternatively, both problems can be solved by the rank-dependent or cumulative functional, first proposed by Quiggin (1982) for decision under risk and by Schmeidler (1989) for decision under uncertainty. Instead of transforming each probability separately, this model transforms the entire cumulative distribution function. The present theory applies the cumulative functional separately to gains and to losses. This development extends prospect theory to uncertain as well as to risky prospects with any number of outcomes while preserving most of its essential features. The differences between the cumulative and the original versions of the theory are discussed in section "Relation to previous work".

Let S be a finite set of states of nature; subsets of S are called events. It is assumed that exactly one state obtains, which is unknown to the decision maker. Let X be a set of consequences, also called outcomes. For simplicity, we confine the present

discussion to monetary outcomes. We assume that X includes a neutral outcome, denoted 0, and we interpret all other elements of X as gains or losses, denoted by positive or negative numbers, respectively.

An uncertain prospect f is a function from S into X that assigns to each state $s \in S$ a consequence $f(s) = x$ in X. To define the cumulative functional, we arrange the outcomes of each prospect in increasing order. A prospect f is then represented as a sequence of pairs (x_i, A_i), which yields x_i if A_i occurs, where $x_i > x_j$ iff $i > j$, and (A_i) is a partition of S. We use positive subscripts to denote positive outcomes, negative subscripts to denote negative outcomes, and the zero subscript to index the neutral outcome. A prospect is called strictly positive or positive, respectively, if its outcomes are all positive or nonnegative. Strictly negative and negative prospects are defined similarly; all other prospects are called mixed. The positive part of f denoted f^+, is obtained by letting $f^+(s) = f(s)$ if $f(s) > 0$, and $f^+(s) = 0$ if $f(s) \leq 0$. The negative part of f, denoted f^-, is defined similarly.

As in expected utility theory, we assign to each prospect f a number $V(f)$ such that f is preferred to or indifferent to g iff $V(f) \geq V(g)$. The following representation is defined in terms of the concept of *capacity* (Choquet 1955), a nonadditive set function that generalizes the standard notion of probability. A capacity W is a function that assigns to each $A \subset S$ a number $W(A)$ satisfying $W(\emptyset) = 0$, $W(S) = 1$, and $W(A) \geq W(B)$ whenever $A \supset B$.

Cumulative prospect theory asserts that there exist a strictly increasing value function v: $X \to \mathrm{Re}$, satisfying $v(x_0) = v(0) = 0$, and capacities W^+ and W^-, such that for $f = (x_i, A_i)$, $-m \leq i \leq n$,

$$V(f) = V\left(f^+\right) + V\left(f^-\right),$$
$$V\left(f^+\right) = \sum_{i=0}^{n} \pi_i^+ v(x_i), \quad V\left(f^-\right) = \sum_{i=-m}^{0} \pi_i^- v(x_i), \tag{24.1}$$

where the decision weights $\pi^+\left(f^+\right) = \left(\pi_0^+, \ldots, \pi_n^+\right)$ and $\pi^-\left(f^-\right) = \left(\pi_{-m}^-, \ldots, \pi_0^-\right)$ are defined by:

$$\pi_n^+ = W^+(A_n), \pi_{-m}^- = W^-(A_{-m}),$$
$$\pi_i^+ = W^+(A_i \cup \cdots \cup A_n) - W^+(A_{i+1} \cup \cdots \cup A_n), 0 \leq i \leq n-1,$$
$$\pi_i^- = W^-(A_{-m} \cup \cdots \cup A_i) - W^-(A_{-m} \cup \cdots \cup A_{i-1}), 1-m \leq i \leq 0.$$

Letting $\pi_i = \pi_i^+$ if $i \geq 0$ and $\pi_i = \pi_i^-$ if $i < 0$, Eq. (24.1) reduces to

$$V(f) = \sum_{i=-m}^{n} \pi_i v(x_i). \tag{24.2}$$

The decision weight π_i^+, associated with a positive outcome, is the difference between the capacities of the events "the outcome is at least as good as x_i" and "the outcome is strictly better then x_i." The decision weight π_i^-, associated with

a negative outcome, is the difference between the capacities of the events "the outcome is at least as bad as x_i" and "the outcome is strictly worse than x_i." Thus, the decision weight associated with an outcome can be interpreted as the marginal contribution of the respective event,[1] defined in terms of the capacities W^+ and W^-. If each W is additive, and hence a probability measure, then π_i is simply the probability of A_i. It follows readily from the definitions of π and W that for both positive and negative prospects, the decision weights add to 1. For mixed prospects, however, the sum can be either smaller or greater than 1, because the decision weights for gains and for losses are defined by separate capacities.

If the prospect $f = (x_i, A_i)$ is given by a probability distribution $p(A_i) = p_i$, it can be viewed as a probabilistic or risky prospect (x_i, p_i). In this case, decision weights are defined by:

$$\begin{aligned}
&\pi_n^+ = w^+(p_n), \pi_{-m}^- = w^-(p_{-m}),\\
&\pi_i^+ = w^+(p_i + \cdots + p_n) - w^+(p_{i+1} + \cdots + p_n), 0 \le i \le n-1,\\
&\pi_i^- = w^-(p_{-m} + \cdots + p_i) - w^-(p_{-m} + \cdots + p_{i-1}), 1-m \le i \le 0.
\end{aligned}$$

where w^+ and w^- are strictly increasing functions from the unit interval into itself satisfying $w^+(0) = w^-(0) = 0$, and $w^+(1) = w^-(1) = 1$.

To illustrate the model, consider the following game of chance. You roll a die once and observe the result $x = 1, \ldots, 6$. If x is even, you receive \$$x$; if x is odd, you pay \$$x$. Viewed as a probabilistic prospect with equiprobable outcomes, f yields the consequences $(-5, -3, -1, 2, 4, 6)$, each with probability $1/6$. Thus, $f^+ = (0, 1/2; 2, 1/6; 4, 1/6; 6, 1/6)$, and $f^- = (-5, 1/6; -3, 1/6; -1, 1/6; 0, 1/2)$. By Eq. (24.1), therefore,

$$\begin{aligned}
V(f) &= V(f^+) + V(f^-)\\
&= v(2)\left[w^+(1/2) - w^+(1/3)\right] + v(4)\left[w^+(1/3) - w^+(1/6)\right]\\
&\quad + v(6)\left[w^+(1/6) - w^+(0)\right]\\
&\quad + v(-5)\left[w^-(1/6) - w^-(0)\right] + v(-3)\left[w^-(1/3) - w^-(1/6)\right]\\
&\quad + v(-1)\left[w^-(1/2) - w^-(1/3)\right].
\end{aligned}$$

Relation to Previous Work

Luce and Fishburn (1991) derived essentially the same representation from a more elaborate theory involving an operation $\circ$ of joint receipt or multiple play. Thus, $f \circ g$ is the composite prospect obtained by playing both f and g, separately. The key feature of their theory is that the utility function U is additive with respect to $\circ$ that is, $U(f \circ g) = U(f) + U(g)$ provided one prospect is acceptable (i.e.,

[1] In keeping with the spirit of prospect theory, we use the decumulative form for gains and the cumulative form for losses. This notation is vindicated by the experimental findings described in section "Experiment".

preferred to the status quo) and the other is not. This condition seems too restrictive both normatively and descriptively. As noted by the authors, it implies that the utility of money is a linear function of money if for all sums of money x, y, $U(x \bigcirc y) = U(x + y)$. This assumption appears to us inescapable because the joint receipt of x and y is tantamount to receiving their sum. Thus, we expect the decision maker to be indifferent between receiving a $10 bill or receiving a $20 bill and returning $10 in change. The Luce-Fishburn theory, therefore, differs from ours in two essential respects. First, it extends to composite prospects that are not treated in the present theory. Second, it practically forces utility to be proportional to money.

The present representation encompasses several previous theories that employ the same decision weights for all outcomes. Starmer and Sugden (1989) considered a model in which $w^-(p) = w^+(p)$, as in the original version of prospect theory. In contrast, the rank-dependent models assume $w^-(p) = 1 - \mathrm{w}^+(1-p)$ or $W^-(A) = 1 - W^+(S-A)$. If we apply the latter condition to choice between uncertain assets, we obtain the choice model established by Schmeidler (1989), which is based on the Choquet integral.[2] Other axiomatizations of this model were developed by Gilboa (1987), Nakamura (1990), and Wakker (1989a, b). For probabilistic (rather than uncertain) prospects, this model was first established by Quiggin (1982) and Yaari (1987), and was further analyzed by Chew (1989), Segal (1989), and Wakker (1990). An earlier axiomatization of this model in the context of income inequality was presented by Weymark (1981). Note that in the present theory, the overall value $V(f)$ of a mixed prospect is not a Choquet integral but rather a sum $V(f^+) + V(f^-)$ of two such integrals.

The present treatment extends the original version of prospect theory in several respects. First, it applies to any finite prospect and it can be extended to continuous distributions. Second, it applies to both probabilistic and uncertain prospects and can, therefore, accommodate some form of source dependence. Third, the present theory allows different decision weights for gains and losses, thereby generalizing the original version that assumes $w^+ = w^-$. Under this assumption, the present theory coincides with the original version for all two-outcome prospects and for all mixed three-outcome prospects. It is noteworthy that for prospects of the form $(x,p;y,\ 1-p)$, where either $x > y > 0$ or $x < y < 0$, the original theory is in fact rank dependent. Although the two models yield similar predictions in general, the cumulative version—unlike the original one—satisfies stochastic dominance. Thus, it is no longer necessary to assume that transparently dominated prospects are eliminated in the editing phase—an assumption that was criticized by some authors. On the other hand, the present version can no longer explain violations of stochastic dominance in nontransparent contexts (e.g., Tversky and Kahneman 1986). An axiomatic analysis of the present theory and its relation to cumulative utility theory and to expected utility theory are discussed in the appendix; a more comprehensive treatment is presented in Wakker and Tversky (1991).

[2]This model appears under different names. We use *cumulative utility theory* to describe the application of a Choquet integral to a standard utility function, and *cumulative prospect theory* to describe the application of two separate Choquet integrals to the value of gains and losses.

Values and Weights

In expected utility theory, risk aversion and risk seeking are determined solely by the utility function. In the present theory, as in other cumulative models, risk aversion and risk seeking are determined jointly by the value function and by the capacities, which in the present context are called cumulative weighting functions, or weighting functions for short. As in the original version of prospect theory, we assume that v is concave above the reference point ($v''(x) \leq 0$, $x \geq 0$) and convex below the reference point ($v''(x) \geq 0$, $x \leq 0$). We also assume that v is steeper for losses than for gains $v'(x) < v'(-x)$ for $x \geq 0$. The first two conditions reflect the principle of diminishing sensitivity: the impact of a change diminishes with the distance from the reference point. The last condition is implied by the principle of loss aversion according to which losses loom larger than corresponding gains (Tversky and Kahneman 1991).

The principle of diminishing sensitivity applies to the weighting functions as well. In the evaluation of outcomes, the reference point serves as a boundary that distinguishes gains from losses. In the evaluation of uncertainty, there are two natural boundaries—certainty and impossibility—that correspond to the endpoints of the certainty scale. Diminishing sensitivity entails that the impact of a given change in probability diminishes with its distance from the boundary. For example, an increase of .1 in the probability of winning a given prize has more impact when it changes the probability of winning from .9 to 1.0 or from 0 to .1, than when it changes the probability of winning from .3 to .4 or from .6 to .7. Diminishing sensitivity, therefore, gives rise to a weighting function that is concave near 0 and convex near 1. For uncertain prospects, this principle yields subadditivity for very unlikely events and superadditivity near certainty. However, the function is not well-behaved near the endpoints, and very small probabilities can be either greatly overweighted or neglected altogether.

Before we turn to the main experiment, we wish to relate the observed non-linearity of preferences to the shape of the weighting function. For this purpose, we devised a new demonstration of the common consequence effect in decisions involving uncertainty rather than risk. Table 24.1 displays a pair of decision problems (I and II) presented in that order to a group of 156 money managers during a workshop. The participants chose between prospects whose outcomes were contingent on the difference d between the closing values of the Dow-Jones today and tomorrow. For example, f' pays \$25,000 if d exceeds 30 and nothing otherwise. The percentage of respondents who chose each prospect is given in brackets. The independence axiom of expected utility theory implies that f is preferred to g iff f' is preferred to g'. Table 24.1 shows that the modal choice was f in problem I and g' in problem II. This pattern, which violates independence, was chosen by 53 % of the respondents.

Essentially the same pattern was observed in a second study following the same design. A group of 98 Stanford students chose between prospects whose outcomes were contingent on the point-spread d in the forthcoming Stanford-Berkeley football game. Table 24.2 presents the prospects in question. For example, g pays \$10 if

Table 24.1 A test of independence (Dow-Jones)

		A	B	C	
		If $d < 30$	If $30 \leq d \leq 35$	If $35 < d$	
Problem I:	f	$25,000	$25,000	$25,000	[68]
	g	$25,000	0	$75,000	[32]
Problem II:	f'	0	$25,000	$25,000	[23]
	g'	0	0	$75,000	[77]

Note: Outcomes are contingent on the difference d between the closing values of the Dow-Jones today and tomorrow. The percentage of respondents ($N = 156$) who selected each prospect is given in brackets

Table 24.2 A test of independence (Stanford-Berkeley football game)

		A	B	C	
		If $d < 0$	If $0 \leq d \leq 10$	If $10 < d$	
Problem I:	f	$10	$10	$10	[64]
	g	$10	$30	0	[36]
Problem II:	f'	0	$10	$10	[34]
	g'	0	$30	0	[66]

Note: Outcomes are contingent on the point-spread d in a Stanford-Berkeley football game. The percentage of respondents ($N = 98$) who selected each prospect is given in brackets

Stanford does not win, $30 if it wins by 10 points or less, and nothing if it wins by more than 10 points. Ten percent of the participants, selected at random, were actually paid according to one of their choices. The modal choice, selected by 46 % of the subjects, was f and g', again in direct violation of the independence axiom.

To explore the constraints imposed by this pattern, let us apply the present theory to the modal choices in Table 24.1, using $1,000 as a unit. Since f is preferred to g in problem I,

$$v(25) > v(75)W^{+}(C) + v(25)\left[W^{+}(A \cup C) - W^{+}(C)\right]$$

or

$$v(25)\left[1 - W^{+}(A \cup C) + W^{+}(C)\right] > v(75)W^{+}(C).$$

The preference for g' over f' in problem II, however, implies

$$v(75)W^{+}(C) > v(25)W^{+}(C \cup B);$$

hence,

$$W^{+}(S) - W^{+}(S - B) > W^{+}(C \cup B) - W^{+}(C). \tag{24.3}$$

Thus, "subtracting" B from certainty has more impact than "subtracting" B from $C \cup B$. Let $W_{+}(D) = 1 - W^{+}(S - D)$, and $w_{+}(p) = 1 - w^{+}(1 - p)$. It follows readily

that Eq. (24.3) is equivalent to the subadditivity of W_+, that is, $W_+(B) + W_+(D) \geq W + (B \cup D)$. For probabilistic prospects, Eq. (24.3) reduces to

$$1 - w^+(1-q) > w^+(p+q) - w^+(p),$$

or

$$w_+(q) + w_+(r) \geq w_+(q+r), q + r < 1.$$

Allais's example corresponds to the case where $p(C) = .10$, $p(B) = .89$, and $p(A) = .01$.

It is noteworthy that the violations of independence reported in Tables 24.1 and 24.2 are also inconsistent with regret theory, advanced by Loomes and Sugden (1982a, b), and with Fishburn's (1988) SSA model. Regret theory explains Allais's example by assuming that the decision maker evaluates the consequences as if the two prospects in each choice are statistically independent. When the prospects in question are defined by the same set of events, as in Tables 24.1 and 24.2, regret theory (like Fishburn's SSA model) implies independence, since it is additive over states. The finding that the common consequence effect is very much in evidence in the present problems undermines the interpretation of Allais's example in terms of regret theory.

The common consequence effect implies the subadditivity of W_+ and of w_+. Other violations of expected utility theory imply the subadditivity of W^+ and of w^+ for small and moderate probabilities. For example, Prelec (1990) observed that most respondents prefer 2% to win \$20,000 over 1% to win \$30,000; they also prefer 1% to win \$30,000 and 32% to win \$20,000 over 34% to win \$20,000. In terms of the present theory, these data imply that $w^+(.02) - w^+(.01) \geq w^+(.34) - w^+(.33)$. More generally, we hypothesize

$$w^+(p+q) - w^+(q) \geq w^+(p+q+r) - w^+(q+r), \tag{24.4}$$

provided $p + q + r$ is sufficiently small. Equation (24.4) states that w^+ is concave near the origin; and the conjunction of the above inequalities implies that, in accord with diminishing sensitivity, w^+ has an inverted S-shape: it is steepest near the endpoints and shallower in the middle of the range. For other treatments of decision weights, see Hogarth and Einhorn (1990), Prelec (1989), Viscusi (1989), and Wakker (1990). Experimental evidence is presented in the next section.

Experiment

An experiment was carried out to obtain detailed information about the value and weighting functions. We made a special effort to obtain high-quality data. To this end, we recruited 25 graduate students from Berkeley and Stanford (12 men and

13 women) with no special training in decision theory. Each subject participated in three separate one-hour sessions that were several days apart. Each subject was paid $25 for participation.

Procedure

The experiment was conducted on a computer. On a typical trial, the computer displayed a prospect (e.g., 25 % chance to win $150 and 75 % chance to win $50) and its expected value. The display also included a descending series of seven sure outcomes (gains or losses) logarithmically spaced between the extreme outcomes of the prospect. The subject indicated a preference between each of the seven sure outcomes and the risky prospect. To obtain a more refined estimate of the certainty equivalent, a new set of seven sure outcomes was then shown, linearly spaced between a value 25 % higher than the lowest amount accepted in the first set and a value 25 % lower than the highest amount rejected. The certainty equivalent of a prospect was estimated by the midpoint between the lowest accepted value and the highest rejected value in the second set of choices. We wish to emphasize that although the analysis is based on certainty equivalents, the data consisted of a series of choices between a given prospect and several sure outcomes. Thus, the cash equivalent of a prospect was derived from observed choices, rather than assessed by the subject. The computer monitored the internal consistency of the responses to each prospect and rejected errors, such as the acceptance of a cash amount lower than one previously rejected. Errors caused the original statement of the problem to reappear on the screen.[3]

The present analysis focuses on a set of two-outcome prospects with monetary outcomes and numerical probabilities. Other data involving more complicated prospects, including prospects defined by uncertain events, will be reported elsewhere. There were 28 positive and 28 negative prospects. Six of the prospects (three nonnegative and three nonpositive) were repeated on different sessions to obtain the estimate of the consistency of choice. Table 24.3 displays the prospects and the median cash equivalents of the 25 subjects.

A modified procedure was used in eight additional problems. In four of these problems, the subjects made choices regarding the acceptability of a set of mixed prospects (e.g., 50 % chance to lose $100 and 50 % chance to win x) in which x was systematically varied. In four other problems, the subjects compared a fixed prospect (e.g., 50 % chance to lose $20 and 50 % chance to win $50) to a set of prospects (e.g., 50 % chance to lose $50 and 50 % chance to win x) in which x was systematically varied. (These prospects are presented in Table 24.6.)

[3] An IBM disk containing the exact instructions, the format, and the complete experimental procedure can be obtained from the authors.

Table 24.3 Median cash equivalents (in dollars) for all nonmixed prospects

	Probability								
Outcomes	.01	.05	.10	.25	.50	.75	.90	.95	.99
(0, 50)			9		21		37		
(0, –50)			−8		−21		−39		
(0, 100)		14		25	36	52		78	
(0, –100)		−8		−23.5	−42	−63		−84	
(0, 200)	10		20		76		131		188
(0, –200)	−3		−23		−89		−155		−190
(0, 400)	12								377
(0, –400)	−14								−380
(50, 100)			59		71		83		
(–50, –100)			−59		−71		−85		
(50, 150)		64		72.5	86	102		128	
(–50, –150)		−60		−71	−92	−113		−132	
(100, 200)		118		130	141	162		178	
(–100, –200)		−112		−121	−142	−158		−179	

Note: The two outcomes of each prospect are given in the left-hand side of each row; the probability of the second (i.e., more extreme) outcome is given by the corresponding column. For example, the value of $9 in the upper left corner is the median cash equivalent of the prospect (0, .9; $50, .1)

Results

The most distinctive implication of prospect theory is the fourfold pattern of risk attitudes. For the nonmixed prospects used in the present study, the shapes of the value and the weighting functions imply risk-averse and risk-seeking preferences, respectively, for gains and for losses of moderate or high probability. Furthermore, the shape of the weighting functions favors risk seeking for small probabilities of gains and risk aversion for small probabilities of loss, provided the outcomes are not extreme. Note, however, that prospect theory does not imply perfect reflection in the sense that the preference between any two positive prospects is reversed when gains are replaced by losses. Table 24.4 presents, for each subject, the percentage of risk-seeking choices (where the certainty equivalent exceeded expected value) for gains and for losses with low ($p \leq .1$) and with high ($p \geq .5$) probabilities. Table 24.4 shows that for $p \geq .5$, all 25 subjects are predominantly risk averse for positive prospects and risk seeking for negative ones. Moreover, the entire fourfold pattern is observed for 22 of the 25 subjects, with some variability at the level of individual choices.

Although the overall pattern of preferences is clear, the individual data, of course, reveal both noise and individual differences. The correlations, across subjects, between the cash equivalents for the same prospects on successive sessions averaged .55 over six different prospects. Table 24.5 presents means (after transformation to Fisher's z) of the correlations between the different types of prospects. For example, there were 19 and 17 prospects, respectively, with high probability of gain and high

Table 24.4 Percentage of risk-seeking choices

	Gain		Loss	
Subject	$p \leq .1$	$p \geq .5$	$p \leq .1$	$p \geq .5$
1	100	38	30	100
2	85	33	20	75
3	100	10	0	93
4	71	0	30	58
5	83	0	20	100
6	100	5	0	100
7	100	10	30	86
8	87	0	10	100
9	16	0	80	100
10	83	0	0	93
11	100	26	0	100
12	100	16	10	100
13	87	0	10	94
14	100	21	30	100
15	66	0	30	100
16	60	5	10	100
17	100	15	20	100
18	100	22	10	93
19	60	10	60	63
20	100	5	0	81
21	100	0	0	100
22	100	0	0	92
23	100	31	0	100
24	71	0	80	100
25	100	0	10	87
Risk seeking	78[a]	10	20	87[a]
Risk neutral	12	2	0	7
Risk averse	10	88[a]	80[a]	6

Note: The percentage of risk-seeking choices is given for low ($p \leq .1$) and high ($p \geq .5$) probabilities of gain and loss for each subject (risk-neutral choices were excluded). The overall percentage of risk-seeking, risk-neutral, and risk-averse choices for each type of prospect appear at the bottom of the table

[a]Values that correspond to the fourfold pattern

probability of loss. The value of .06 in Table 24.5 is the mean of the $17 \times 19 = 323$ correlations between the cash equivalents of these prospects.

The correlations between responses within each of the four types of prospects average .41, slightly lower than the correlations between separate responses to the same problems. The two negative values in Table 24.5 indicate that those subjects who were more risk averse in one domain tended to be more risk seeking in the other. Although the individual correlations are fairly low, the trend is consistent: 78 % of the 403 correlations in these two cells are negative. There is also a tendency for

Table 24.5 Average correlations between certainty equivalents in four types of prospects

	L^+	H^+	L^-	H^-
L^+	.41	.17	−.23	.05
H^+		.39	.05	−.18
L^-			.40	.06
H^-				.44

Note: Low probability of gain = L^+; high probability of gain = H^+; low probability of loss = L^-; high probability of loss = H^-

subjects who are more risk averse for high-probability gains to be less risk seeking for gains of low probability. This trend, which is absent in the negative domain, could reflect individual differences either in the elevation of the weighting function or in the curvature of the value function for gains. The very low correlations in the two remaining cells of Table 24.5, averaging .05, indicate that there is no general trait of risk aversion or risk seeking. Because individual choices are quite noisy, aggregation of problems is necessary for the analysis of individual differences.

The fourfold pattern of risk attitudes emerges as a major empirical generalization about choice under risk. It has been observed in several experiments (see, e.g., Cohen et al. 1987), including a study of experienced oil executives involving significant, albeit hypothetical, gains and losses (Wehrung 1989). It should be noted that prospect theory implies the pattern demonstrated in Table 24.4 within the data of individual subjects, but it does not imply high correlations across subjects because the values of gains and of losses can vary independently. The failure to appreciate this point and the limited reliability of individual responses has led some previous authors (e.g., Hershey and Schoemaker 1980) to underestimate the robustness of the fourfold pattern.

Scaling

Having established the fourfold pattern in ordinal and correlational analyses, we now turn to a quantitative description of the data. For each prospect of the form $(x, p; 0,1 - p)$, let c/x be the ratio of the certainty equivalent of the prospect to the nonzero outcome x. Figures 24.1 and 24.2 plot the median value of c/x as a function of p, for positive and for negative prospects, respectively. We denote c/x by a circle if $|x| < 200$, and by a triangle if $|x| \geq 200$. The only exceptions are the two extreme probabilities (.01 and .99) where a circle is used for $|x| = 200$. To interpret Figs. 24.1 and 24.2, note that if subjects are risk neutral, the points will lie on the diagonal; if subjects are risk averse, all points will lie below the diagonal in Fig. 24.1 and above the diagonal in Fig. 24.2. Finally, the triangles and the circles will lie on top of each other if preferences are homogeneous, so that multiplying the outcomes of a prospect f by a constant $k > 0$ multiplies its cash equivalent $c(kf)$ by the same

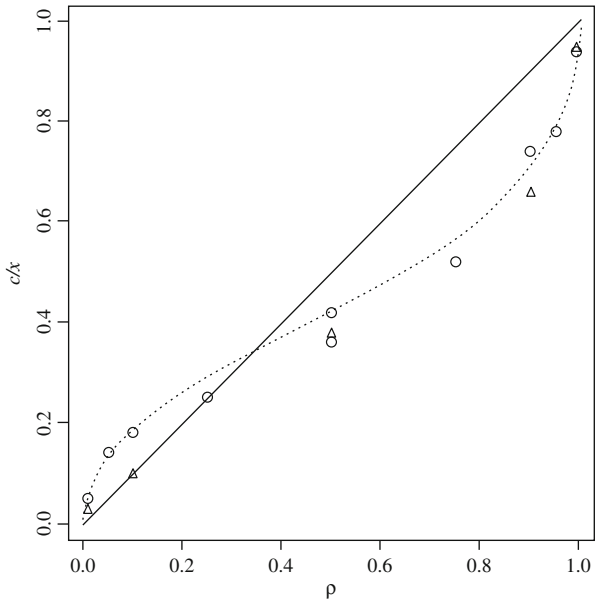


Fig. 24.1 Median c/x for all positive prospects of the form $(x, p; 0{,}1 - p)$. Triangles and circles, respectively, correspond to values of x that lie above or below 200

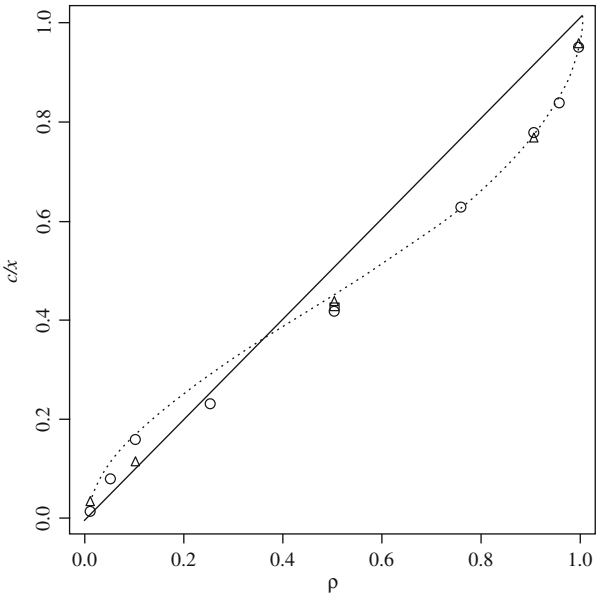


Fig. 24.2 Median c/x for all negative prospects of the form $(x, p; 0{,}1 - p)$. Triangles and circles, respectively, correspond to values of x that lie below or above – 200

constant, that is, $c(kf) = kc(f)$. In expected utility theory, preference homogeneity gives rise to constant relative risk aversion. Under the present theory, assuming $X = \mathrm{Re}$, preference homogeneity is both necessary and sufficient to represent v as a two-part power function of the form

$$v(x) = \begin{cases} x^{\alpha} \text{if } x \geq 0 \\ -\lambda(-x)^{\beta} \text{if } x < 0. \end{cases} \tag{24.5}$$

Figures 24.1 and 24.2 exhibit the characteristic pattern of risk aversion and risk seeking observed in Table 24.4. They also indicate that preference homogeneity holds as a good approximation. The slight departures from homogeneity in Fig. 24.1 suggest that the cash equivalents of positive prospects increase more slowly than the stakes (triangles tend to lie below the circles), but no such tendency is evident in Fig. 24.2. Overall, it appears that the present data can be approximated by a two-part power function. The smooth curves in Figs. 24.1 and 24.2 can be interpreted as weighting functions, assuming a linear value function. They were fitted using the following functional form:

$$w^{+}(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}}, \quad w^{-}(p) = \frac{p^{\delta}}{(p^{\delta} + (1-p)^{\delta})^{1/\delta}}. \tag{24.6}$$

This form has several useful features: it has only one parameter; it encompasses weighting functions with both concave and convex regions; it does not require $w(.5) = .5$; and most important, it provides a reasonably good approximation to both the aggregate and the individual data for probabilities in the range between .05 and .95.

Further information about the properties of the value function can be derived from the data presented in Table 24.6. The adjustments of mixed prospects to acceptability (problems 1–4) indicate that, for even chances to win and lose, a prospect will only be acceptable if the gain is at least twice as large as the loss. This observation is compatible with a value function that changes slope abruptly at zero, with a loss-aversion coefficient of about 2 (Tversky and Kahneman 1991). The median matches in problems 5 and 6 are also consistent with this estimate: when the possible loss is increased by k the compensating gain must be increased by about $2k$. Problems 7 and 8 are obtained from problems 5 and 6, respectively, by positive translations that turn mixed prospects into strictly positive ones. In contrast to the large values of θ observed in problems 1–6, the responses in problems 7 and 8 indicate that the curvature of the value function for gains is slight. A decrease in the smallest gain of a strictly positive prospect is fully compensated by a slightly larger increase in the largest gain. The standard rank-dependent model, which lacks the notion of a reference point, cannot account for the dramatic effects of small translations of prospects illustrated in Table 24.6.

The estimation of a complex choice model, such as cumulative prospect theory, is problematic. If the functions associated with the theory are not constrained, the

Table 24.6 A test of loss aversion

Problem	a	b	c	x	θ
1	0	0	−25	61	2.44
2	0	0	−50	101	2.02
3	0	0	−100	202	2.02
4	0	0	−150	280	1.87
5	−20	50	−50	112	2.07
6	−50	150	−125	301	2.01
7	50	120	20	149	0.97
8	100	300	25	401	1.35

Note: In each problem, subjects determined the value of x that makes the prospect ($a, ½, $b, ½) as attractive as ($c, ½, $x, ½). The median values of x are presented for all problems along with the fixed values a, b, c. The statistic $\theta = (x-b)/(c-a)$ is the ratio of the "slopes" at a higher and a lower region of the value function

number of estimated parameters for each subject is too large. To reduce this number, it is common to assume a parametric form (e.g., a power utility function), but this approach confounds the general test of the theory with that of the specific parametric form. For this reason, we focused here on the qualitative properties of the data rather than on parameter estimates and measures of fit. However, in order to obtain a parsimonious description of the present data, we used a nonlinear regression procedure to estimate the parameters of Eqs. (24.5) and (24.6), separately for each subject. The median exponent of the value function was 0.88 for both gains and losses, in accord with diminishing sensitivity. The median λ was 2.25, indicating pronounced loss aversion, and the median values of γ and δ, respectively, were 0.61 and 0.69, in agreement with Eqs. (24.3) and (24.4) above.[4] The parameters estimated from the median data were essentially the same. Figure 24.3 plots w^+ and w^- using the median estimates of γ and δ.

Figure 24.3 shows that, for both positive and negative prospects, people overweight low probabilities and underweight moderate and high probabilities. As a consequence, people are relatively insensitive to probability difference in the middle of the range. Figure 24.3 also shows that the weighting functions for gains and for losses are quite close, although the former is slightly more curved than the latter (i.e., $\gamma < \delta$). Accordingly, risk aversion for gains is more pronounced than risk seeking for losses, for moderate and high probabilities (see Table 24.3). It is noteworthy that the condition $w^+(p) = w^-(p)$, assumed in the original version of prospect theory, accounts for the present data better than the assumption $w^+(p) = 1 - w^-(1-p)$,

[4]Camerer and Ho (1991) applied Eq. (24.6) to several studies of risky choice and estimated γ from aggregate choice probabilities using a logistic distribution function. Their mean estimate (.56) was quite close to ours.

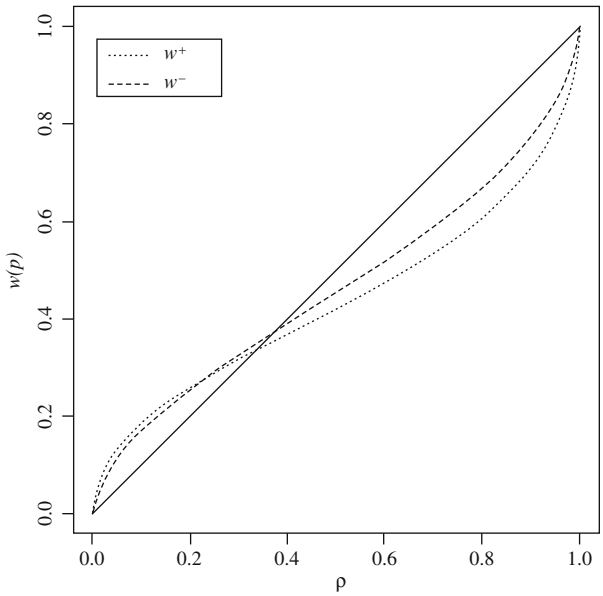


Fig. 24.3 Weighting functions for gains (w^+) and for losses (w^-) based on median estimates of γ and δ in Eq. (24.6)

implied by the standard rank-dependent or cumulative functional. For example, our estimates of w^+ and w^- show that all 25 subjects satisfied the conditions w^+ $(.5) < .5$ and w^- $(.5) < .5$, implied by the former model, and no one satisfied the condition w^+ $(.5) < .5$ iff w^- $(.5) > .5$, implied by the latter model.

Much research on choice between risky prospects has utilized the triangle diagram (Marschak 1950; Machina 1987) that represents the set of all prospects of the form $(x_1,\ p_1;x_2,p_2;x_3,p_3)$, with fixed outcomes $x_1 < x_2 < x_3$. Each point in the triangle represents a prospect that yields the lowest outcome (x_1) with probability p_1, the highest outcome (x_3) with probability p_3, and the intermediate outcome (x_2) with probability $p_2 = 1 - p_1 - p_3$. An indifference curve is a set of prospects (i.e., points) that the decision maker finds equally attractive. Alternative choice theories are characterized by the shapes of their indifference curves. In particular, the indifference curves of expected utility theory are parallel straight lines. Figures 24.4a, b illustrate the indifference curves of cumulative prospect theory for nonnegative and nonpositive prospects, respectively. The shapes of the curves are determined by the weighting functions of Fig. 24.3; the values of the outcomes (x_1, x_2, x_3) merely control the slope.

Figures 24.4a, b are in general agreement with the main empirical generalizations that have emerged from the studies of the triangle diagram; see Camerer (1992), and Camerer and Ho (1991) for reviews. First, departures from linearity, which violate expected utility theory, are most pronounced near the edges of the triangle. Second, the indifference curves exhibit both fanning in and fanning out. Third, the curves are concave in the upper part of the triangle and convex in the lower right.

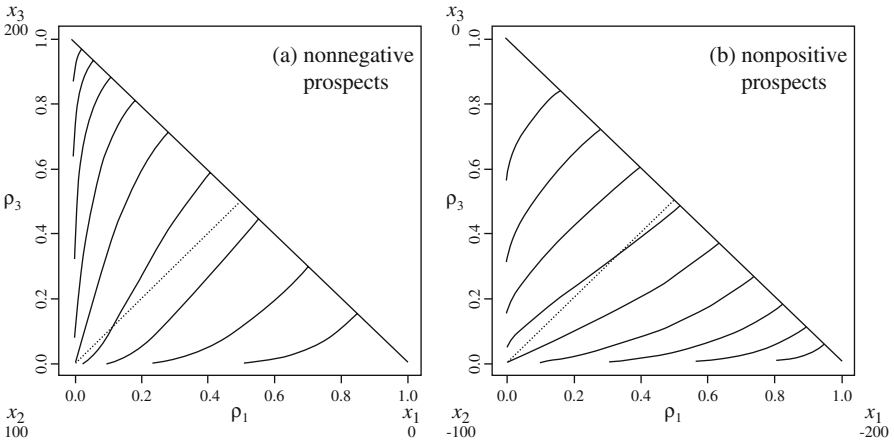


Fig. 24.4 Indifference curves of cumulative prospect theory (**a**) nonnegative prospects ($x_1 = 0$, $x_2 = 100$, $x_3 = 200$), and (**b**) nonpositive prospects ($x_1 = -200$, $x_2 = -100$, $x_3 = 0$). The curves are based on the respective weighting functions of Fig. 24.3, ($\gamma = .61$, $\delta = .69$) and on the median estimates of the exponents of the value function ($\alpha = \beta = .88$). The broken line through the origin represents the prospects whose expected value is x_2

Finally, the indifference curves for nonpositive prospects resemble the curves for nonnegative prospects reflected around the 45° line, which represents risk neutrality. For example, a sure gain of $100 is equally as attractive as a 71 % chance to win $200 or nothing (see Fig. 24.4a), and a sure loss of $100 is equally as aversive as a 64 % chance to lose $200 or nothing (see Fig. 24.4b). The approximate reflection of the curves is of special interest because it distinguishes the present theory from the standard rank-dependent model in which the two sets of curves are essentially the same.

Incentives

We conclude this section with a brief discussion of the role of monetary incentives. In the present study we did not pay subjects on the basis of their choices because in our experience with choice between prospects of the type used in the present study, we did not find much difference between subjects who were paid a flat fee and subjects whose payoffs were contingent on their decisions. The same conclusion was obtained by Camerer (1989), who investigated the effects of incentives using several hundred subjects. He found that subjects who actually played the gamble gave essentially the same responses as subjects who did not play; he also found no differences in reliability and roughly the same decision time. Although some studies found differences between paid and unpaid subjects in choice between simple prospects, these differences were not large enough to change any significant qualitative conclusions. Indeed, all major violations of expected utility theory

(e.g. the common consequence effect, the common ratio effect, source dependence, loss aversion, and preference reversals) were obtained both with and without monetary incentives.

As noted by several authors, however, the financial incentives provided in choice experiments are generally small relative to people's incomes. What happens when the stakes correspond to three- or four-digit rather than one- or two-digit figures? To answer this question, Kachelmeier and Shehata (1991) conducted a series of experiments using Masters students at Beijing University, most of whom had taken at least one course in economics or business. Due to the economic conditions in China, the investigators were able to offer subjects very large rewards. In the high payoff condition, subjects earned about three times their normal monthly income in the course of one experimental session! On each trial, subjects were presented with a simple bet that offered a specified probability to win a given prize, and nothing otherwise. Subjects were instructed to state their cash equivalent for each bet. An incentive compatible procedure (the BDM scheme) was used to determine, on each trial, whether the subject would play the bet or receive the "official" selling price. If departures from the standard theory are due to the mental cost associated with decision making and the absence of proper incentives, as suggested by Smith and Walker (1992), then the highly paid Chinese subjects should not exhibit the characteristic nonlineariry observed in hypothetical choices, or in choices with small payoffs.

However, the main finding of Kachelmeier and Shehata (1991) is massive risk seeking for small probabilities. Risk seeking was slightly more pronounced for lower payoffs, but even in the highest payoff condition, the cash equivalent for a 5 % bet (their lowest probability level) was, on average, three times larger than its expected value. Note that in the present study the median cash equivalent of a 5 % chance to win \$100 (see Table 24.3) was \$14, almost three times the expected value of the bet. In general, the cash equivalents obtained by Kachelmeier and Shehata were higher than those observed in the present study. This is consistent with the finding that minimal selling prices are generally higher than certainty equivalents derived from choice (see, e.g., Tversky et al. 1990). As a consequence, they found little risk aversion for moderate and high probability of winning. This was true for the Chinese subjects, at both high and low payoffs, as well as for Canadian subjects, who either played for low stakes or did not receive any payoff. The most striking result in all groups was the marked overweighting of small probabilities, in accord with the present analysis.

Evidently, high incentives do not always dominate noneconomic considerations, and the observed departures from expected utility theory cannot be rationalized in terms of the cost of thinking. We agree with Smith and Walker (1992) that monetary incentives could improve performance under certain conditions by eliminating careless errors. However, we maintain that monetary incentives are neither necessary nor sufficient to ensure subjects' cooperativeness, thoughtfulness, or truthfulness. The similarity between the results obtained with and without monetary incentives in choice between simple prospects provides no special reason for skepticism about experiments without contingent payment.

Discussion

Theories of choice under uncertainty commonly specify (1) the objects of choice, (2) a valuation rule, and (3) the characteristics of the functions that map uncertain events and possible outcomes into their subjective counterparts. In standard applications of expected utility theory, the objects of choice are probability distributions over wealth, the valuation rule is expected utility, and utility is a concave function of wealth. The empirical evidence reported here and elsewhere requires major revisions of all three elements. We have proposed an alternative descriptive theory in which (1) the objects of choice are prospects framed in terms of gains and losses, (2) the valuation rule is a two-part cumulative functional, and (3) the value function is S-shaped and the weighting functions are inverse S-shaped. The experimental findings confirmed the qualitative properties of these scales, which can be approximated by a (two-part) power value function and by identical weighting functions for gains and losses.

The curvature of the weighting function explains the characteristic reflection pattern of attitudes to risky prospects. Overweighting of small probabilities contributes to the popularity of both lotteries and insurance. Underweighting of high probabilities contributes both to the prevalence of risk aversion in choices between probable gains and sure things, and to the prevalence of risk seeking in choices between probable and sure losses. Risk aversion for gains and risk seeking for losses are further enhanced by the curvature of the value function in the two domains. The pronounced asymmetry of the value function, which we have labeled loss aversion, explains the extreme reluctance to accept mixed prospects. The shape of the weighting function explains the certainty effect and violations of quasi-convexity. It also explains why these phenomena are most readily observed at the two ends of the probability scale, where the curvature of the weighting function is most pronounced (Camerer 1992).

The new demonstrations of the common consequence effect, described in Tables 24.1 and 24.2, show that choice under uncertainty exhibits some of the main characteristics observed in choice under risk. On the other hand, there are indications that the decision weights associated with uncertain and with risky prospects differ in important ways. First, there is abundant evidence that subjective judgments of probability do not conform to the rules of probability theory (Kahneman et al. 1982). Second, Ellsberg's example and more recent studies of choice under uncertainty indicate that people prefer some sources of uncertainty over others. For example, Heath and Tversky (1991) found that individuals consistently preferred bets on uncertain events in their area of expertise over matched bets on chance devices, although the former are ambiguous and the latter are not. The presence of systematic preferences for some sources of uncertainty calls for different weighting functions for different domains, and suggests that some of these functions lie entirely above others. The investigation of decision weights for uncertain events emerges as a promising domain for future research.

The present theory retains the major features of the original version of prospect theory and introduces a (two-part) cumulative functional, which provides a con-

venient mathematical representation of decision weights. It also relaxes some descriptively inappropriate constraints of expected utility theory. Despite its greater generality, the cumulative functional is unlikely to be accurate in detail. We suspect that decision weights may be sensitive to the formulation of the prospects, as well as to the number, the spacing and the level of outcomes. In particular, there is some evidence to suggest that the curvature of the weighting function is more pronounced when the outcomes are widely spaced (Camerer 1992). The present theory can be generalized to accommodate such effects, but it is questionable whether the gain in descriptive validity, achieved by giving up the separability of values and weights, would justify the loss of predictive power and the cost of increased complexity.

Theories of choice are at best approximate and incomplete. One reason for this pessimistic assessment is that choice is a constructive and contingent process. When faced with a complex problem, people employ a variety of heuristic procedures in order to simplify the representation and the evaluation of prospects. These procedures include computational shortcuts and editing operations, such as eliminating common components and discarding nonessential differences (Tversky 1969). The heuristics of choice do not readily lend themselves to formal analysis because their application depends on the formulation of the problem, the method of elicitation, and the context of choice.

Prospect theory departs from the tradition that assumes the rationality of economic agents; it is proposed as a descriptive, not a normative, theory. The idealized assumption of rationality in economic theory is commonly justified on two grounds: the conviction that only rational behavior can survive in a competitive environment, and the fear that any treatment that abandons rationality will be chaotic and intractable. Both arguments are questionable. First, the evidence indicates that people can spend a lifetime in a competitive environment without acquiring a general ability to avoid framing effects or to apply linear decision weights. Second, and perhaps more important, the evidence indicates that human choices are orderly, although not always rational in the traditional sense of this word.

Appendix: Axiomatic Analysis

Let $F = \{f : S \to X\}$ be the set of all prospects under study, and let F^+ and F^- denote the positive and the negative prospects, respectively. Let $\precsim$ be a binary preference relation on F, and let $\approx$ and $>$ denote its symmetric and asymmetric parts, respectively. We assume that $\precsim$ is complete, transitive, and strictly monotonic, that is, if $f \neq g$ and $f(s) \geq g(s)$ for all $s \in S$, then $f > g$.

For any $f, g \in F$ and $A \subset S$, define $h = fAg$ by: $h(s) = f(s)$ if $s \in A$, and $h(s) = g(s)$ if $s \in S - A$. Thus, fAg coincides with f on A and with g on $S - A$. A preference relation $\precsim$ on F satisfies *independence* if for all $f, g, f', g' \in F$ and $A \subset S$, $fAg \precsim fAg'$ iff $f' Ag \precsim f' Ag'$. This axiom, also called the sure thing principle (Savage 1954), is one of the basic qualitative properties underlying expected utility theory, and it is violated by Allais's common consequence effect. Indeed, the attempt to accommodate Allais's example has motivated the development of numerous

models, including cumulative utility theory. The key concept in the axiomatic analysis of that theory is the relation of comonotonicity, due to Schmeidler (1989). A pair of prospects $f, g \in F$ are *comonotonic* if there are no $s, t \in S$ such that $f(s) > f(t)$ and $g(t) > g(s)$. Note that a constant prospect that yields the same outcome in every state is comonotonic with all prospects. Obviously, comonotonicity is symmetric but not transitive.

Cumulative utility theory does not satisfy independence in general, but it implies independence whenever the prospects $fAg, fAg', f'Ag$, and $f'Ag'$ above are pairwise comonotonic. This property is called *comonotonic independence*.[5] It also holds in cumulative prospect theory, and it plays an important role in the characterization of this theory, as will be shown below. Cumulative prospect theory satisfies an additional property, called *double matching:* for all $f, g \in F$, if $f^+ \approx g^+$ and $f^- \approx g^-$, then $f \approx g$.

To characterize the present theory, we assume the following structural conditions: S is finite and includes at least three states; $X = \mathrm{Re}$; and the preference order is continuous in the product topology on Re^k, that is, $\{f \in F : f \succsim g\}$ and $\{f \in F : g \succsim f\}$ are closed for any $g \in F$. The latter assumptions can be replaced by restricted solvability and a comonotonic Archimedean axiom (Wakker 1991).

Theorem 24.1 Suppose $(F^+, \succsim)$ and $(F^-, \succsim)$ can each be represented by a cumulative functional. Then $(F, \succsim)$ satisfies cumulative prospect theory iff it satisfies double matching and comonotonic independence.

The proof of the theorem is given at the end of the appendix. It is based on a theorem of Wakker (1992) regarding the additive representation of lower-diagonal structures. Theorem 24.1 provides a generic procedure for characterizing cumulative prospect theory. Take any axiom system that is sufficient to establish an essentially unique cumulative (i.e., rank-dependent) representation. Apply it separately to the preferences between positive prospects and to the preferences between negative prospects, and construct the value function and the decision weights separately for F^+ and for F^-. Theorem 24.1 shows that comonotonic independence and double matching ensure that, under the proper rescaling, the sum $V(f^+) + V(f^-)$ preserves the preference order between mixed prospects. In order to distinguish more sharply between the conditions that give rise to a one-part or a two-part representation, we need to focus on a particular axiomatization of the Choquet functional. We chose Wakker's (1989a, b) because of its generality and compactness.

For $x \in X$, $f \in F$, and $r \in S$, let $x\{r\}f$ be the prospect that yields x in state r and coincides with f in all other states. Following Wakker (1989a), we say that a preference relation satisfies *tradeoff consistency*[6] (TC) if for all $x, x', y, y' \in X, f, f', g, g' \in F$, and $s, t \in S$.

[5] Wakker (1989b) called this axiom *comonotonic coordinate independence*. Schmeidler (1989) used *comonotonic independence* for the mixture space version of this axiom: $f \precsim g$ iff. $\alpha f + (1-\alpha)h \precsim \alpha g + (1-\alpha)h$.

[6] Wakker (1989a, b) called this property *cardinal coordinate independence.* He also introduced an equivalent condition, called the absence of *contradictory tradeoffs.*

$$x\{s\}f \lesssim y\{s\}g, x'\{s\}f \gtrsim y'\{s\}g \text{ and } x\{t\}f' \gtrsim y\{t\}g' \text{ imply } x'\{t\}f' \gtrsim y'\{t\}g'.$$

To appreciate the import of this condition, suppose its premises hold but the conclusion is reversed, that is, $y'\{t\}g' > x'\{t\}f'$. It is easy to verify that under expected utility theory, the first two inequalities, involving $\{s\}$, imply $u(y) - u(y') \geq u(x) - u(x')$, whereas the other two inequalities, involving $\{t\}$, imply the opposite conclusion. Tradeoff consistency, therefore, is needed to ensure that "utility intervals" can be consistently ordered. Essentially the same condition was used by Tversky et al. (1988) in the analysis of preference reversal, and by Tversky and Kahneman (1991) in the characterization of constant loss aversion.

A preference relation satisfies *comonotonic tradeoff consistency* (CTC) if TC holds whenever the prospects $x\{s\}f$, $y\{s\}g$, $x'\{s\}f$, and $y'\{s\}g$ are pairwise comonotonic, as are the prospects $x\{f\}f'$, $y\{t\}g'$, $x'\{t\}f'$, and $y'\{t\}g'$ (Wakker 1989a). Finally, a preference relation satisfies *sign-comonotonic tradeoff consistency* (SCTC) if CTC holds whenever the consequences x, x', y, y' are either all nonnegative or all nonpositive. Clearly, TC is stronger than CTC, which is stronger than SCTC. Indeed, it is not difficult to show that (1) expected utility theory implies TC, (2) cumulative utility theory implies CTC but not TC, and (3) cumulative prospect theory implies SCTC but not CTC. The following theorem shows that, given our other assumptions, these properties are not only necessary but also sufficient to characterize the respective theories.

Theorem 24.2 Assume the structural conditions described above.

(a) (Wakker 1989a) Expected utility theory holds iff $\lesssim$ satisfies TC.
(b) (Wakker 1989b) Cumulative utility theory holds iff $\lesssim$ satisfies CTC.
(c) Cumulative prospect theory holds iff $\lesssim$ satisfies double matching and SCTC.

A proof of part c of the theorem is given at the end of this section. It shows that, in the presence of our structural assumptions and double matching, the restriction of tradeoff consistency to sign-comonotonic prospects yields a representation with a reference-dependent value function and different decision weights for gains and for losses.

Proof of Theorem 24.1 The necessity of comonotonic independence and double matching is straightforward. To establish sufficiency, recall that, by assumption, there exist functions π^+, π^-, v^+, v^-, such that $V^+ = \sum \pi^+ v^+$ and $V^- = \sum \pi^- v^-$ preserve $\lesssim$ on F^+ and on F^-, respectively. Furthermore, by the structural assumptions, π^+ and π^- are unique, whereas v^+ and v^- are continuous ratio scales. Hence, we can set $v^+(1) = 1$ and $v^-(-1) = \theta < 0$, independently of each other.

Let Q be the set of prospects such that for any $q \in Q$, $q(s) \neq q(t)$ for any distinct $s, t \in S$. Let F_g denote the set of all prospects in F that are comonotonic with G. By comonotonic independence and our structural conditions, it follows readily from a theorem of Wakker (1992) on additive representations for lower-triangular subsets of Re^k that, given any $q \in Q$, there exist intervals scales $\{U_{qi}\}$, with a common unit, such that $U_q = \sum_i U_{qi}$ preserves $\lesssim$ on F_q. With no loss of generality we

can set $U_{qi}(0) = 0$ for all i and $U_q(1) = 1$. Since V^+ and V^- above are additive representations of $\lesssim$ on F_q^+ and F_q^-, respectively, it follows by uniqueness that there exist aq, $bq > 0$ such that for all i, U_{qi} equals $a_q \pi_i^+ v^+$ on Re^+, and U_{qi} equals $b_q \pi_i^- v^-$ on Re^-.

So far the representations were required to preserve the order only within each F_q. Thus, we can choose scales so that $b_q = 1$ for all q. To relate the different representations, select a prospect $h \neq q$. Since V^+ should preserve the order on F^+, and U_q should preserve the order within each F_q, we can multiply V^+ by a_h, and replace each a_q by a_q/a_h. In other words, we may set $a_h = 1$. For any $q \in Q$, select $f \in F_q$, $g \in F_h$ such that $f^+ \approx \mathrm{g}^+ > 0$, $f^- \approx g^- > 0$, and $\mathrm{g} \approx 0$. By double matching, then, $f \approx g \approx 0$. Thus, $a_q V^+ (f^+) + V^-(f^-) = 0$, since this form preserves the order on F_q. But $V^+(f^+) = V^+(g^+)$ and $\mathrm{V}^-(f^-) = V^-(g^-)$, so $V^+(g^+) + V^-(g^-) = 0$ implies $V^+(f^+) + V^-(f^-) = 0$. Hence, $a_q = 1$, and $V(f) = V^+(f^+) + V^-(f^-)$ preserves the order within each F_q.

To show that V preserves the order on the entire set, consider any f, $g \in F$ and suppose $f \lesssim g$. By transitivity, $c(f) \lesssim c(g)$ where $c(f)$ is the certainty equivalent of f. Because $c(f)$ and $c(g)$ are comonotonic, $V(f) = V(c(f)) \geq V(c(g)) = V(g)$. Analogously, $f > g$ implies $V(f) > V(g)$, which complete the proof of theorem 24.1.

Proof of Theorem 24.2 (part c) To establish the necessity of SCTC, apply cumulative prospect theory to the hypotheses of SCTC to obtain the following inequalities:

$$\begin{aligned} V(x\{s\}f) &= \pi_s v(x) + \sum_{r \in S-s} \pi_r v(f(r)) \\ &\leq \pi'_s v(y) + \sum_{r \in S-s} \pi'_r v(g(r)) = V(y\{s\}g) \\ V(x'\{s\}f) &= \pi_s v(x') + \sum_{r \in S-s} \pi_r v(f(r)) \\ &\geq \pi'_s v(y') + \sum_{r \in S-s} \pi'_r v(g(r)) = V\left(y'\{s\}g\right). \end{aligned}$$

The decision weights above are derived, assuming SCTC, in accord with Eqs. (24.1) and (24.2). We use primes to distinguish the decision weights associated with g from those associated with f. However, all the above prospects belong to the same comonotonic set. Hence, two outcomes that have the same sign and are associated with the same state have the same decision weight. In particular, the weights associated with $x\{s\}f$ and $x'\{s\}f$ are identical, as are the weights associated with $y\{s\}g$ and with $y'\{s\}g$. These assumptions are implicit in the present notation. It follows that

$$\pi_s v(x) - \pi'_s v(y) \leq \pi_s v\left(x'\right) - \pi'_s v\left(y'\right).$$

Because x, y, x', y' have the same sign, all the decision weights associated with state s are identical, that is, $\pi_s = \pi'_s$. Cancelling this common factor and rearranging terms yields $v(y) - v(y') \geq v(x) - v(x')$.

Suppose SCTC is not valid, that is, $x\{t\}f \lesssim y\{t\}g'$ but $x'\{t\}f' < y'\{t\}g'$. Applying cumulative prospect theory, we obtain

$$\begin{aligned} V(x\{t\}f') &= \pi_t v(x) + \sum_{r\in S-t} \pi_r v\left(f'(r)\right) \\ &\geq \pi_t v(y) + \sum_{r\in S-t} \pi_r v\left(g'(r)\right) = V(y\{t\}g') \\ V(x'\{t\}f') &= \pi_t v(x') + \sum_{r\in S-t} \pi_r v\left(f'(r)\right) \\ &< \pi_t v(y') + \sum_{r\in S-t} \pi_t v\left(g'(r)\right) = V(y'\{t\}g'). \end{aligned}$$

Adding these inequalities yields $v(x) - v(x') > v(y) - v(y')$ contrary to the previous conclusion, which establishes the necessity of SCTC. The necessity of double matching is immediate.

To prove sufficiency, note that SCTC implies comonotonic independence. Letting $x = y$, $x' = y'$, and $f = g$ in TC yields $x\{t\}f' \lesssim x\{t\}g'$ implies $x'\{t\}f' \lesssim x'\{t\}g'$, provided all the above prospects are pairwise comonotonic. This condition readily entails comonotonic independence (see Wakker 1989b).

To complete the proof, note that SCTC coincides with CTC on $(F^+, \lesssim)$ and on $(F^-, \lesssim)$. By part b of this theorem, the cumulative functional holds, separately, in the nonnegative and in the nonpositive domains. Hence, by double matching and comonotonic independence, cumulative prospect theory follows from Theorem 24.1.

References

Allais, M. (1953). Le comportement de l'homme rationel devant le risque, critique des postulates et axiomes de l'ecole americaine. *Econometrica, 21*, 503–546.

Arrow, K. J. (1982). Risk perception in psychology and economies. *Economic Inquiry, 20*, 1–9.

Camerer, C. F. (1989). An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty, 2*, 61–104.

Camerer, C. F. (1992). Recent tests of generalizations of expected utility theory. In W. Edwards (Ed.), *Utility: Theories, measurement and applications*. Boston: Kluwer Academic Publishers.

Camerer, C. F., & Ho, T.-H., (1991). Nonlinear weighting of probabilities and violations of the betweenness axiom. Unpublished manuscript, The Wharton School, University of Pennsylvania.

Chew, S.-H. (1989). An axiomatic generalization of the quasilinear mean and the gini mean with application to decision theory, Unpublished manuscript, Department of Economics, University of California at Irvine.

Choquet, G. (1955). Theory of capacities. *Annales de L'Institut Fourier, 5*, 131–295.

Cohen, M., Jaffray, J.-Y., & Said, T. (1987). Experimental comparison of individual behavior under risk and under uncertainty for gains and for losses. *Organizational Behavior and Human Decision Processes, 39*, 1–22.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics, 75*, 643–669.

Fishburn, P. C. (1988). *Nonlinear preference and utility theory*. Baltimore: The Johns Hopkins University Press.

Gilboa, I. (1987). Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics, 16*, 65–88.

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty, 4*, 5–28.

Hershey, J. C., & Schoemaker, P. J. H. (1980). Prospect theory's reflection hypothesis: A critical examination. *Organizational Behavior and Human Performance, 25*, 395–418.

Hogarth, R., & Einhorn, H. (1990). Venture theory: A model of decision weights. *Management Science, 36*, 780–803.

Kachelmeier, S. J., & Shehata, M. (1992). Examining risk preferences under high monetary incentives: Experimental evidence from the People's Republic of China. *American Economic Review, 82*(5), 1120–1141.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.

Kahneman, D., & Tversky, A. (1984). Choices, values and frames. *American Psychologist, 39*, 341–350.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Loomes, G., & Sugden, R. (1982a). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal, 92*, 805–824.

Loomes, G., & Sugden, R. (1982b). Some implications of a more general form of regret theory. *Journal of Economic Theory, 41*, 270–287.

Luce, R. D., & Fishburn, P. C. (1991). Rank- and sign-dependent linear utility models for finite first-order gambles. *Journal of Risk and Uncertainty, 4*, 29–59.

Machina, M. J. (1987). Choice under uncertainty: Problems solved and unsolved. *Economic Perspectives, 1*(1), 121–154.

Marschak, J. (1950). Rational behavior, uncertain prospects, and measurable utility. *Econometrica, 18*, 111–114.

Nakamura, Y. (1990). Subjective expected utility with non-additive probabilities on finite state space. *Journal of Economic Theory, 51*, 346–366.

Prelec, D. (1989). On the shape of the decision weight function. Unpublished manuscript, Harvard Graduate School of Business Administration.

Prelec, D. (1990). A 'Pseudo-endowment' effect, and its implications for some recent non-expected utility models. *Journal of Risk and Uncertainty, 3*, 247–259.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization, 3*, 323–343.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica, 57*, 571–587.

Segal, U. (1989). Axiomatic representation of expected utility with rank-dependent probabilities. *Annals of Operations Research, 19*, 359–373.

Smith, V. L., & Walker, J. M. (1992). Monetary rewards and decision cost in experimental economics. Unpublished manuscript, Economic Science Lab, University of Arizona.

Starmer, C., & Sugden, R. (1989). Violations of the independence axiom in common ratio problems: An experimental test of some competing hypotheses. *Annals of Operations Research, 19*, 79–102.

Tversky, A. (1969). The intransitivity of preferences. *Psychology Review, 76*, 31–48.

Tversky, A., Kahneman, D. (1986). Rational choice and the framing of decisions, *The Journal of Business 59*(4), part 2, S251–S278.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference dependent model. *Quarterly Journal of Economics, 107*(4), 1039–1061.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95*(3), 371–384.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *The American Economic Review, 80*(1), 204–217.

Viscusi, K. W. (1989). Prospective reference theory: Toward an explanation of the paradoxes. *Journal of Risk and Uncertainty, 2*, 235–264.

Wakker, P. P. (1989a). *Additive representations of preferences: A new foundation in decision analysis*. Dordrecht: Kluwer Academic Publishers.

Wakker, P. P. (1989b). Continuous subjective expected utility with nonadditive probabilities. *Journal of Mathematical Economics, 18*, 1–27.

Wakker, P. P. (1990). Separating marginal utility and risk aversion. Unpublished manuscript, University of Nijmegen, The Netherlands.

Wakker, P. P. (1991). Additive representations of preferences, a new foundation of decision analysis; the algebraic approach. In J. D. Doignon & J. C. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp. 71–87). Berlin: Springer.

Wakker, P. (1993). Additive representations on rank-ordered sets: II. The topological approach. *Journal of Mathematical Economics, 22*(1), 1–26.

Wakker, P. P., & Tversky, A. (1991). An axiomatization of cumulative prospect theory. Unpublished manuscript. University of Nijmegan, the Netherlands.

Wehrung, D. A. (1989). Risk taking over gains and losses: A study of oil executives. *Annals of Operations Research, 19*, 115–139.

Weymark, J. A. (1981). Generalized gini inequality indices. *Mathematical Social Sciences, 1*, 409–430.

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica, 55*, 95–115.

Part IV
Logics of Knowledge and Belief

Chapter 25
Introduction

Horacio Arló-Costa, Vincent F. Hendricks, and Johan van Benthem

This part, like all others in this book, consists of a mix of classic papers that have defined the area and modern ones illustrating important current issues. These texts provide rich fare, and they defy simple labels summarizing their content. Moreover, a look at the list of authors reveals a mixture of different academic cultures, from philosophy to computer science. One might also add that the texts do not all agree: they span a landscape with many positions and perspectives.

Epistemic logic as the systematic study of reasoning with knowledge and belief started with Jaakko Hintikka's classic book Knowledge and Belief: An Introduction to Logic of the Two Notions, which set the agenda for many subsequent lines of research and debate. It interpreted knowledge as what is true in some current range of epistemically accessible worlds, while doing something similar for belief and doxastic accessibility. Thus, general methods from modal logic became available for studying knowledge, and the resulting axiomatic systems have shaped many philosophical discussions for or against principles of 'omniscience', 'closure', and positive and negative 'introspection'. One general interest behind such specific issues has been the search for satisfactory definitions of knowledge, an interest with a long epistemological history running from Plato's "justified true belief" to post-Gettier strengthenings involving forms of robustness of true beliefs under new information, under new relevant considerations, or across counterfactual variations

H. Arló-Costa (deceased)
Carnegie Mellon University, Pittsburgh, PA, USA

V.F. Hendricks (✉)
Center for Information and Bubble Studies, University of Copenhagen, Copenhagen, Denmark
e-mail: vincent@hum.ku.dk

J. van Benthem
University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Stanford University, Stanford, United States
e-mail: johan@science.uva.nl

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_25

of the actual world – and in a related spirit, in various forms of faithful tracking during a history of investigation. The past half-century of formal studies has even produced further perspectives, such as the availability of proof or evidence in some precise sense, or convergence in the limit in some process of inquiry.

The paper by Dretske makes the original Hintikka semantics a more dynamic process, showing how knowledge claims are always based on some current range of relevant worlds, which can change under the pressure of legitimate new considerations. In a related vein, Lewis provides systematic principles guiding this process of selecting relevant worlds. Following a further intuition, Nozick proposes a counterfactual idea of knowledge as a true belief that would stay attuned to the facts in non-actual worlds close to ours. While these approaches are semantic, Artemov's 'justification logic' brings explicit proof and evidence into epistemic logic, allowing us to syntactically manipulate reasons for our beliefs. Finally, Kelly discusses the learning-theoretic view of temporal convergence for knowledge of complete histories of the world.

Since this part is about logics of knowledge and belief, many readers will be interested not just in formal languages and semantics, but also in complete calculi for reasoning capturing the operational proof-theoretic aspects of reasoning with knowledge or belief. Stalnaker's paper presents a broad logical view of possible modal systems and defensible epistemic and doxastic principles, and in another perspective, so does Artemov. Interestingly, not all notions of knowledge proposed in the 'robustness' tradition have been studied in this systematic manner, and many questions remain open, though actively pursued by some young philosophical logicians.

Another running theme is the issue of which epistemic attitudes form a natural family that requires scrutiny in its entirety. Knowledge and belief need not be enough, and for instance, Stalnaker's survey of doxastic and epistemic logics proposes a new notion of 'safe belief' that will survive true new information, as an intermediate between logic and belief simpliciter. Parikh even suggests an algebraic framework that ties together a wide abstract range of knowledge, belief, and action in general. Specializing general action again to informational action, we arrive at what has been Hintikka's guiding interest throughout: the combination of knowledge as based on our current semantic information with acts that systematically change that information, such as questions, and games of inquiry over time.

It is only a short step then to a dynamic focus on learning rather than the mere statics of knowledge. Kelly's article looks at this dynamics from the viewpoint of learning theory, and investigates which truths can be acquired in the limit, i.e., which processes of inquiry will reliably converge to stable true belief about the answer to the main question at stake. In a somewhat related mode, Williamson focuses on the scope of purely operational definitions of knowledge, and shows that knowledge-based epistemology remains indispensable. And these are just two dynamic or computational aspects of knowledge and belief. There is more to information dynamics when one begins to study effects of specific local acts of

knowledge update or belief revision as an agent navigates the world. Many of these topics will be addressed in the next section on interactive epistemology, since much information flow that is crucial to humans involves more than one party: be it a group of agents, or just one agent interacting with Nature.

There are many further themes to ponder when reading these articles. Does the semantics describe internal first-person views of epistemic agents, or the theorist's external view of their situation? Do different epistemic attitudes correlate with different sorts of information? How does knowledge of propositions tie in with knowledge of objects, "that" or "whether" versus "which", and why not then also discuss knowledge "how" and "why"? And finally, what is the status of all these logical theories? Are they normative prescriptions, of do they represent some existing cognitive practice, if only idealized and at higher levels of abstraction? Reading the papers in this section will not necessarily answer all these questions, but it will make readers much better equipped to pursue these issues for themselves.

Suggested Further Reading

Starting with a classical trailblazer, J. Hintikka, Knowledge and Belief: An Introduction to the Logic of the Two Notions, Cornell University Press 1962 and King's College Publications 2005, set the whole subject on its course. A series of later books broadened the paradigm to a general view of information and inquiry, as represented in J. Hintikka, Logic, Language-Games and Information: Kantian Themes in the Philosophy of Logic, Clarendon Press Oxford, 1973. Putting inquiry at center stage in epistemology has also been the persistent theme of Robert Stalnaker's work in the field, with Inquiry, The MIT Press, 1987, as a classic source. Meanwhile, a richer view of possible epistemic and doxastic attitudes suggested by natural language was investigated in W. Lenzen, "Recent Work in Epistemic Logic", Acta Philosophica Fennica 30 (1978): 1–219, which also discusses links with probability. Also still in the 1970s, epistemic and doxastic logic were rediscovered in the foundations of game theory, but references for this will be found in another part of these readings. But arguably the major development invigorating epistemic logic has been its crossing into computer science, in the study of information-driven agency. Two major books demonstrating the resulting outburst of new research are R. Fagin, J. Y. Halpern, Y. Moses & M. Vardi, Reasoning about Knowledge, The MIT Press, 1995, and W. van der Hoek & J-J Meijer, Epistemic Logic for AI and Computer Science, Cambridge University Press, 1995. An even more radically computational algorithmic understanding has been that of formal learning theory, inspired by learning methods for infinite structures like human languages or the process of scientific inquiry. The classic source for this is K. Kelly, The Logic of Reliable Inquiry, Oxford University Press, 1996. We conclude with two other angles on knowledge that bring in further mathematical paradigms. One is the verificationist perspective on knowledge through proof and evidence, for which a classic text is M. Dummett, Truth and Other Enigmas, Harvard University Press, 1978. Finally, while these publications are concerned with knowledge and belief, another broad stream has taken information flow through Shannon-type channels to be the basic underlying notion, following Dretske's classic Knowledge and the Flow of Information, The MIT Press, 1981. An innovative logical framework taking this road much further is J. Barwise & J. Seligman, Information Flow, Cambridge University Press, 1995.

Chapter 26
Epistemology Without Knowledge and Without Belief

Jaakko Hintikka

Knowledge and Decision-Making

Epistemology seems to enjoy an unexpectedly glamorous reputation in these days. A few years ago, William Safire (1995) wrote a popular novel called *The Sleeper Spy*. It depicts a distinctly post-Cold War world in which it is no longer easy to tell the good guys—including the good spies—from the bad ones. To emphasize this sea change, Safire tells us that his Russian protagonist has not been trained in the military or in the police, as he would have been in the old days, but as an epistemologist.

But is this with-it image deserved? Would the theory of knowledge that contemporary academic epistemologists cultivate be of any help to a sleeper spy? This question prompts a critical survey of the state of the art or, rather, the state of the theory of knowledge. I submit that the up-to-date image is not accurate and that most of the current epistemological literature deals with unproductive and antiquated questions. This failure is reflected in the concepts that are employed by contemporary epistemologists.

What are those concepts? It is usually thought and said that the most central concepts of epistemology are knowledge and belief. The prominence of these two notions is reflected in the existing literature on epistemology. A large chunk of it consists in discussions of how the concept of knowledge is to be defined or is not to be defined. Are those discussions on the target? An adequate analysis of such concepts as knowledge and belief, whether it is calculated to lead us to a formal

Jaakko Hintikka was deceased at the time of publication.

J. Hintikka (deceased)
Boston University, Helsinki, Finland

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_26

definition or not, should start from the role that they play in real life. Now in real life we are both producers and consumers of knowledge. We acquire knowledge in whatever ways we do so, and we then put it to use in our actions and decision-making. I will here start from the latter role, which takes us to the question: What is the role that the notion of knowledge plays in that decision-making?

To take a simple example, let us suppose that I am getting ready to face a new day in the morning. How, then, does it affect my actions if I know that it will not rain today? You will not be surprised if I say that what it means is that I am entitled to behave as if it will not rain—for instance to leave my umbrella home. However, you may be surprised if I claim that most of the important features of the logical behavior of the notion of knowledge can be teased out of such simple examples. Yet this is the case. My modest example can be generalized. The role of knowledge in decision-making is to rule out certain possibilities. In order to use my knowledge, I must know which possibilities it rules out. In other words, any one scenario must therefore be either incompatible or compatible with what I know, for I am either entitled or not entitled to disregard it. Thus the totality of incompatible scenarios determines what I know and what I do not know, and vice versa. In principle, all that there is to logic of knowledge is this dichotomy between epistemically impossible and epistemically possible scenarios.

It is also clear how this dichotomy serves the purposes of decision-making, just as it does in my mini-example of deciding whether or not to take an umbrella with me. But the connection with overt behavior is indirect, for what the dichotomy merely demarcates are the limits of what I am *entitled to* disregard. And being entitled to do something does not always mean that I do it. It does not always show up in the overt ways one actually or even potentially acts. For other considerations may very well enter into my decision-making. Maybe I just want to sport an umbrella even though I know that it need not serve its function of shielding myself from rain. Maybe I am an epistemological *akrates* and act against what I know. The connection is nevertheless real, even though it is a subtle one. There is a link between my knowledge and my decisions, but it is, so to speak, a *de jure* connection and not a *de facto* connection. I think that this is a part of what John Austin (1961a) was getting at when he compared "I know" with "I promise." To know something does not mean simply to have evidence of a superior degree for it, nor does it mean to have a superior kind of confidence in it. If my first names were George Edward, I might use the open-question argument to defend these distinctions. By saying "I promise," I entitle you to expect that I fulfill my promise. By saying "I know," I claim that I am entitled to disregard those possibilities that do not agree with what I know. There is an evaluative element involved in the concept of knowledge that does not reduce to the observable facts of the case. Hence, it is already seen to be unlikely that you could define what it means to know by reference to matters of fact, such as the evidence that the putative knower possesses or the state of the knower's mind.

This evaluative element is due to the role of knowledge in guiding our life in that it plays a role in the justification of our decisions. This role determines in the last analysis the logic and in some sense the meaning of knowledge. A Wittgensteinean might put this point by saying that decision-making is one of the language-games

that constitute the logical home of the concept of knowledge. You can remove knowledge from the contexts of decision-making, but you cannot remove a relation to decision-making from the concept of knowledge. For this reason, it is among other things misguided in a fundamental way to try to separate epistemic possibility from actual (natural) possibility. Of course, the two are different notions, but the notion of epistemic possibility has conceptual links to the kind of possibility that we have to heed in our decision-making. For one thing, the set of scenarios involved in the two notions must be the same.

But the main point here is not that there is an evaluative component to the notion of knowledge. The basic insight is that there is a link between the concept of knowledge and human action. The evaluative element is merely a complicating factor in the equation. The existence of a link between the two is not peculiar to the notion of knowledge. There is a link, albeit of a different kind, also in the case of belief. In fact, the conceptual connection is even more obvious in the case of belief. Behavioral scientists have studied extensively decision principles where belief constitutes one component, as, for instance, in the principle of maximizing expected utility. It usually comes in the form of degrees of belief. (They are often identified with probabilities.) Typically, utilities constitute another component. Whether or not such explicit decision principles capture the precise links between belief and behavior, they illustrate the existence of the link and yield clues to its nature.

Indeed, from a systematic point of view, the relative roles assigned to knowledge and to belief in recent epistemology and recent decision theory cannot but appear paradoxical. Belief is in such studies generally thought of as a direct determinant of our decisions, whereas knowledge is related to action only indirectly, if at all. Yet common sense tells us that one of the main reasons for looking for more knowledge is to put us in a better position in our decision-making, whereas philosophers often consider belief—especially when it is contrasted with knowledge—as being initially undetermined by our factual information and therefore being a much worse guide to decision-making. Probability is sometimes said to be a guide to life, but surely knowledge is a better one. Or, if we cannot use black-or-white concepts here, shouldn't rational decision-making be guided by degrees of knowledge rather than degrees of mere belief?

The same point can perhaps be made by noting that in many studies of decision-making, a rational agent is supposed to base his or her decisions on the agent's beliefs (plus, of course, utilities) and then by asking: Would it not be even more rational for the agent to base his or her decisions on what the agent *knows*?

In order for a rational agent to act on his or her belief, this belief clearly must be backed up by some evidence. Otherwise, current decision theory makes little sense. The difference is that the criteria of what entities are to act are different in the case of belief from what they are in the case of knowledge. If I act on a belief, that belief must satisfy my personal requirements for that role. They may vary from person to person. In contrast, the criteria of knowing are impersonal and not dependent on the agent in question. In order to define knowledge as distinguished from beliefs, we would have to spell out those impersonal criteria. This is obviously an extremely difficult task at best.

Another fact that complicates the connection between knowledge and behavior—that is, between what I know and what I do—is that in principle, this link is holistic. What matters to my decisions in the last analysis is the connection between the totality of my knowledge. There is not always any hard-and-fast connection between particular items of knowledge and my behavior. In principle, the connection is via my entire store of knowledge. This is reflected by the fact emphasized earlier that the dichotomy that determines the logic of knowledge is a distinction between scenarios that are ruled out by the *totality* of what I know and scenarios that are compatible with the *totality* of my knowledge and that I therefore must be prepared for. The same feature of the concept of knowledge also shows up in the requirement of total evidence that is needed in Bayesian inference and which has prompted discussion and criticism there. (See, e.g., Earman 1992.)

To spell out the criteria of the justification involved in the applications of the concept of knowledge is to define what knowledge is as distinguished from other propositional attitudes. Characterizing these conditions is obviously a complicated task. I will return to these criteria later in this chapter.

The Logic of Knowledge and Information

Meanwhile, another dimension of the concept of knowledge is brought out by homely examples of the kind I am indulging in. By this time it should be clear—I hope—that it is extremely hard to specify the kind of entitlement or justification that knowing something amounts to. This difficulty is perhaps sufficiently attested to by the inconclusiveness of the extensive discussions about how to define knowledge that one can find in the literature. (See, e.g., Shope 1983.) But another aspect of this notion is in principle as clear as anything one can hope to find in philosophical analysis (or synthesis). It may be difficult to tell whether a certain propositional attitude amounts to knowledge, belief, opinion or whatnot, but there is typically no difficulty in spelling out the *content* of any one of these attitudes on some particular occasion. Here, the lesson drawn from my rain-and-umbrella example is applicable. It was seen that what someone knows specifies, and is specified by, the class of possible scenarios that are compatible with what he or she knows. And such classes of scenarios or of "possible worlds" can be captured linguistically as the classes of scenarios (alias possible worlds) in which a certain sentence is true. Indeed, for Montague (1974, p. 153) such classes of possible worlds (or, strictly speaking, the characteristic functions of these classes, in the sense of functions from possible worlds to truth-values) *are* propositions. In this way, the content of a propositional attitude can normally be captured verbally. For another instance, for Husserl (1983, sec. 124), the task would be to capture the noematic *Sinn* of an act, which he says can in principle always be accomplished linguistically—that is, in Husserl's terminology, through *Bedeutungen*.

Let us now call the members of the class of scenarios admitted by someone's knowledge that someone's epistemic alternatives. That I know that it will not rain

today means that none of the scenarios under which the wet stuff falls down are among my epistemic alternatives, and likewise for all *knowing that* statements. What the concept of knowledge involves in a purely logical perspective is thus a dichotomy of the space of all possible scenarios into those that are compatible with what I know and those that are incompatible with my knowledge. What was just seen is that this dichotomy is directly conditioned by the role of the notion of knowledge in real life. Now this very dichotomy is virtually all we need in developing an explicit logic of knowledge, better known as epistemic logic. This conceptual parentage is reflected by the usual notation of epistemic logic. In it, the epistemic operator K_a ("a knows that") receives its meaning from the dichotomy between excluded and admitted scenarios, while the sentence within its scope specifies the content of the item of knowledge in question.

Basing epistemic logic on such a dichotomy has been the guiding idea of my work in epistemic logic right from the beginning. I have seen this idea being credited to David Lewis, but I have not seen any uses of it that predate my work.

But here we seem to run into a serious problem in interpreting epistemic logic from the vantage point of a dichotomy of excluded and admitted scenarios. Such an interpretation might seem to exclude "quantifying in"—that is to say, to exclude applications of the knowledge operator to open formulas for them, it would not make any sense to speak of scenarios in which the content of one's knowledge is true or false. Such "quantifying in" is apparently indispensable for the purpose of analyzing the all-important *wh*-constructions with *knows*. For instance, "John *knows* who murdered Roger Ackroyd" apparently must be expressed by

$$(\exists x)\, K_{John}\, (x \text{ murdered Roger Ackroyd}) \tag{26.1}$$

as distinguished from

$$K_{John}\, (\exists x)\, (x \text{ murdered Roger Ackroyd}) \tag{26.2}$$

which says that John knows that someone murdered the victim and hence can serve as the presupposition of the question, "Who murdered Roger Ackroyd?"

But in (26.1), the notion of knowledge apparently cannot be interpreted by reference to a distinction between admitted and excluded scenarios. The reason is that the knowledge operator in (26.1) is prefixed to an open formula. Such an open formula cannot be said to be true or false in a given scenario, for its truth depends on the value of the variable x. Hence it cannot implement the required dichotomy.

In order for our epistemic discourse to express the *wh*-constructions, the knowledge operator must apparently be allowed to occur also internally, prefixed to open formulas rather than sentences (formulas without free variables). This prompts a serious interpretational problem. Indeed we can see here the reason for the deep theoretical interest of the problem of "quantifying in," which otherwise might strike one as being merely the logicians' technical problem. Fortunately, this apparent problem can be solved by means of suitable analysis of the relations between different logical operators (see section "Information acquisition as a questioning procedure").

An epistemic logic of this kind can obviously be developed within the framework of possible worlds semantics. (For a sketch of how this can be done, see Hintikka 2003b.) In fact, the truth condition for *knows that* is little more than a translation of what was just said: "b knows that S" is true in a world W if and only if S is true in all the epistemic b-alternatives to W. These alternatives are all the scenarios or "worlds" compatible with everything b knows in W. In certain important ways, this truth condition for knowledge statements is clearer than its counterpart in the ordinary (alethic) modal semantics, in that in epistemic logic the interpretation of the alternativeness relation (alias accessibility relation) is much clearer than in the logic of physical or metaphysical modalities.

Here we have already reached a major conclusion. Epistemic logic presupposes essentially only the dichotomy between epistemically possible and epistemically excluded scenarios. How this dichotomy is drawn is a question pertaining to the definition of knowledge. However, we do not need to know this definition in doing epistemic logic. Thus the logic and the semantics of knowledge can be understood independently of any explicit definition of knowledge. Hence it should not be surprising to see that a similar semantics and a similar logic can be developed for other epistemic notions—for instance, belief, information, memory, and even perception. This is an instance of a general law holding for propositional attitudes. This law says that the content of a propositional attitude can be specified independently of differences between different attitudes. This law has been widely recognized, even if it has not always been formulated as a separate assumption. For instance, in Husserl (1983, e.g., sec. 133) it takes the form of separating the noematic *Sinn* from the thetic component of a noema. As a consequence, the respective logics of different epistemic notions do not differ much from each other. In particular, they do not differ at all in those aspects of their logic that depend merely on the dichotomical character of their semantics. These aspects include prominently the laws that hold for quantifiers and identity, especially the modifications that are needed in epistemic contexts in the laws of the substitutivity of identity and existential generalization.

The fact that different epistemic notions, such as knowledge, belief, and information, share the same dichotomic logic should not be surprising in the light of what has been said. The reason is that they can all serve the same purpose of guiding our decisions, albeit in different ways. Hence the same line of thought can be applied to them as was applied earlier to the concept of knowledge, ending up with the conclusion that their logic is a dichotomic logic not unlike the logic that governs the notion of knowledge. The common ingredient in all these different logics is then the true epistemic logic. But it turns out to be a logic of information rather than a logic of knowledge.

This distinction between what pertains to the mere dichotomy between admitted and excluded scenarios and what pertains to the criteria relied on in this dichotomy is not a novelty. It is at bottom only a restatement in structural terms of familiar contrast, which in the hands of different thinkers has received apparently different formulations. The dichotomy defines the content of a propositional attitude, while the criteria of drawing it determine which propositional attitude we are dealing with.

Hence we are naturally led to the project of developing a generic logic of contents of attitudes, independent of the differences between different attitudes.

This generic logic of epistemology can be thought of as the logic of information. Indeed, what the content of a propositional attitude amounts to can be thought of as a certain item of information. In attributing different attitudes to agents, different things are said about this information—for instance, that it is known, believed, remembered, and so on. This fits in well with the fact that the same content can be known by one person, believed by another, remembered by a third one, and so on. This idea that one and the same objective content may be the target of different people's different attitudes is part of what Frege (see, e.g., 1984) was highlighting by his notion of *the thought*. Thus it might even be happier to talk about the logic of information than about epistemic logic. John Austin (1961b) once excused his use of the term "performative" by saying that even though it is a foreign word and an ugly word that perhaps does not mean very much, it has one good thing about it: It is not a deep word. It seems to me that epistemology would be in much better shape if instead of the deep word "knowledge," philosophers cultivated more the ugly foreign word "information," even though it perhaps does not capture philosophers' profound sense of knowing. In any case, in the generic logic of epistemology here envisaged, philosophers' strong sense of knowledge plays no role.

Information Acquisition as a Questioning Procedure

But what about the other context in which we encounter knowledge in real life—the context of knowledge acquisition? As was noted, what the concept of knowledge amounts to is revealed by two questions: What is it that we are searching for in the process of knowledge acquisition? What purpose can the product of such an inquiry serve? The second question has now been discussed. It remains to examine the crucial first question. Surely the first order of business of any genuine theory of knowledge—the most important task both theoretically and practically—is how new acquired, not merely how previously obtained information can be evaluated. A theory of information (knowledge) acquisition is both philosophically and humanly much more important than a theory of whether or not already achieved information amounts to knowledge. Discovery is more important than the defense of what you already know. In epistemology, as in warfare, offense frequently is the best defense.

This point can be illustrated in a variety of ways. For instance, a thinker who does not acquire any information cannot even be a skeptic, for he or she would not have anything to be skeptical about. And a skeptic's doubts must be grounded on some grasp as to how that information is obtained, unless these doubts are totally irrational. Epistemology cannot start from the experience of wonder or doubt. It should start from recognition of where the item of information that we are wondering about or doubting came from in the first place. Any rational justification or rational distinction of such wonder or doubt must be based on its ancestry.

Fortunately we now have available to us a framework in which to discuss the logic and epistemology of knowledge acquisition or, rather, if I have the terminological courage of my epistemological convictions, information acquisition. The framework is what is referred to as the interrogative model of inquiry or interrogative approach to inquiry. (See Hintikka 1999.) Its basic idea is the same as that of the oldest explicit form of reasoning in philosophy, the Socratic method of questioning or *elenchus*. In it, all new information enters into an argument or a line of reasoning in the form of answers to questions that the inquirer addresses to a suitable source of information.

It might at first seem implausible that this approach might yield a viable theory of ampliative reasoning in general, for several different reasons. Fortunately all these objections can be overcome. First, it might not seem likely that this model can be developed into a form explicit and detailed enough to allow for precise conclusions. This objection would have been eminently appropriate as recently as a decade or two ago. For it is only in the last several years that there has existed a general and explicit logical theory of all the relevant kinds of questions. This logic of questions and answers is the backbone of the interrogative model. This theory has not yet been presented in a monographic or textbook form, but its basic ideas are explained in recent and forthcoming papers of mine. (See, e.g., Hintikka 2003a.) This logic of questions and answers is an extension and application of epistemic logic (logic of knowledge). It has been made possible by a quiet revolution in epistemic logic. One of the main problems in representing questions is to specify which ingredients of the aimed-at information are the questioned elements—that is to say, are supposed to be made known by the answer. It turns out that their specification can sometimes be accomplished only by means of the independence indicators whose logic is only now being explored, even though it cannot be done in the earlier "first-generation" epistemic logic. The details of the new "second-generation" epistemic logic that makes use of the notion of independence need not concern us here. It may nevertheless be noted that this new logic solves the problem of "quantifying in" in that in it, the epistemic operator K always occurs sentence-initially. There is no problem of quantifying in, one might say here, only quantifying (binding variables) independently of an epistemic operator.

Another main requirement that can be addressed to the interrogative approach—and indeed to the theory of any goal-directed activity—is that it must do justice to the strategic aspects of inquiry. Among other things, it ought to be possible to distinguish the definitory rules of the activity in question from its strategic rules. The former spell out what is possible at each stage of the process. The latter express what actions are better and worse for the purpose of reaching the goals of the activity. This requirement can be handled most naturally by doing what Plato already did to the Socratic *elenchus* and by construing knowledge-seeking by questioning as a game that pits the questioner against the answerer. Then the study of the strategies of knowledge acquisition becomes another application of the mathematical theory of games, which perhaps ought to be called "strategy theory" rather than "game theory" in the first place. The distinction between the definitory rules—usually called simply the rules of the game—and strategic principles is built right into the structure of such games.

The greatest obstacle to generality might seem to be the apparently restricted range of applicability of the interrogative model. Some of the resistance to this approach, which I have referred to as the idea of "inquiry as inquiry," can be dispelled by pointing out that questions and answers can be understood in a wide sense, and have to be so understood if the generality claim is to be acceptable. Sources of answers to explicit or implicit questions have to include not only human witnesses and other informants or databases in a computer, but observation and experimentation as well as memory and tacit knowledge. One of the leading ideas of the interrogative approach is that all information used in an argument must be brought in as an answer to a question. In claiming such generality for the interrogative model, I can appeal to such precedents as Collingwood's (1940) and Gadamer's (1975) "logic of questions and answers," even though what they called logic really was not. My claims of generality on behalf of the interrogative approach are not even as sweeping as Collingwood's thesis that every proposition may be considered as an answer to a question. Likewise in construing experiments as questions to nature, I can cite Kant (Kant 1787) and Bacon.

Interrogation and Justification

But the context of knowledge acquisition is vital even if the aim of your game is justification and not discovery. Suppose that a scientist has a reason to think that one of his or her conclusions is not beyond doubt. What is he or she to do? Will the scientist try to mine his or her data so as to extract from them grounds for a decision? Sometimes, perhaps, but in an overwhelming majority of actual scientific situations, the scientist will ask what further information one should in such circumstances try to obtain in order to confirm or disconfirm the suspect proposition—for instance, what experiments it would be advisable to perform or what kinds of observation one should try to make in order to throw light on the subject matter. Unfortunately such contexts—or should I say, such language-games—of verification by means of new information have not received much attention from recent philosophers. They have been preoccupied with the justification of already acquired knowledge rather than with the strategies of reaching new knowledge.

Thus we must extend the scope of the interrogative model in such a way that it enables us to cope with justification and not just pure discovery. What we need is a rule or rules that authorize the rejection—which is tentative and may be only temporary—of some of the answers that an inquirer receives. The *terminus technicus* for such rejection is *bracketing*. The possibility of bracketing widens the scope of epistemological and logical methods tremendously. After this generalization has been carried out, the logic of interrogative inquiry can serve many of the same purposes as the different variants of non-monotonic reasoning, and serve them without the tacit assumptions that often make nonmonotonic reasoning epistemologically restricted or even philosophically dubious. A telling example is offered by what is known as circumscriptive reasoning. (See McCarthy 1990.)

It relies on the assumption that the premises present the reasoner with all the relevant information, so that the reasoner can assume that they are made true in the intended models in the simplest possible way. This is an assumption that in fact can often be made, but it is not always available on other occasions. As every puzzle fan knows, often a key to the clever reasoning needed to solve a puzzle lies precisely in being able to imagine circumstances in which the normal expectations evoked by the specification of the puzzle are not realized. Suppose a puzzle goes as follows: "Evelyn survived George by more than 80 years, even though she was born many decades before him. How come?" The explanation is easy if you disregard the presumption that "George" is a man's name and "Evelyn" a woman's. Evelyn Waugh in fact survived George Eliot by 86 years. Here the solution of the puzzle depends entirely on going beyond the *prima facie* information provided by the putative—in other words, on violating the presuppositions of a circumscriptive inference. Reasoning by circumscription is enthymemic reasoning. It involves tacit premises that may be false.

Thus by introducing the idea of bracketing, we can dispense with all modes of ampliative reasoning. The only rules besides rules of logical inference are the rules for questioning and the rule allowing bracketing. This may at first look like a cheap trick serving merely to sweep all the difficulties of epistemic justification under the rug of bracketing. In reality, what is involved is an important insight. What is involved is not a denial of the difficulties of justification, but an insight into their nature as problems. Once a distinction is made between strategic and definitory rules, it is realized that the definitory rules can only be permissive, telling what one may do in order to reach knowledge and to justify it. The problem of justification is a strategic problem. It pertains to what one ought to do in order to make sure that the results of one's inquiry are secure. This is to be done by the double process of disregarding dubious results and confirming the survivors through further inquiry. The only new permissive rule needed for the purpose is the rule that allows bracketing.

Thus the question as to which answers to bracket is always at bottom a strategic problem. It is therefore futile in principle to try to capture the justificatory process by means of definitory rules of this or that kind. To attempt to do so is a fallacy that in the last analysis vitiates all the usual "logics" of ampliative reasoning. This mistake is committed not only by non-monotonic logics but also by inductive logic and by the current theories of belief revision. Ampliative logics can be of considerable practical interest and value, but in the ultimate epistemological perspective, they are but types of enthymemic reasoning, relying on tacit premises quite as much as circumscriptive reasoning. An epistemologist's primary task here is not to study the technicalities of such modes of reasoning, fascinating though they are in their own right. It is to uncover the tacit premises on which such euthymemic reasoning is in reality predicated.

Allowing bracketing is among other things important because it makes it possible to conceive of interrogative inquiry as a model also of the confirmation of hypotheses and other propositions in the teeth of evidence. The interrogative model can thus also serve as a general model of the justification of hypotheses. It should

in fact be obvious that the processes of discovery and justification cannot be sharply separated from each other in the practice or in the theory of science. Normally, a new discovery in science is justified by the very same process—for instance, by the same experiments—by means of which it was made, or could have been made And this double duty service of questioning is not due only to the practical exigencies of "normal science." It has a firm conceptual basis. This basis is the fact that information (unlike many Federal appropriations) does not come to an inquirer earmarked for a special purpose—for instance, for the purpose of discovery rather than justification. The inquirer may ask a question for this or that proximate purpose in mind, but there is nothing in the answer that rules out its being used for other purposes as well.

And such an answer can only be evaluated in terms of its service for both causes. This is because from game theory we know that in the last analysis, game-like goal-directed processes can be evaluated only in terms of their strategies, not in terms of what one can say of particular moves—for instance, what kinds of "warrants" they might have. As a sports-minded logician might explain the point, evaluating a player's skills in a strategic game is in principle like judging a figure-skating performance rather than keeping score in a football game. In less playful terms, one can in generally associate utilities (payoffs) only with strategies, not with particular moves. But since discovery and justification are aspects of the same process, they have to be evaluated in terms of the different possible strategies that are calculated to serve both purposes.

When we realize this strategic inseparability of the two processes, we can in fact gain a better understanding of certain otherwise puzzling features of epistemic enterprise. For instance, we can now see why it sometimes is appropriate to jump to a conclusion on the basis of relatively thin evidence. The reason is that finding what the truth is can help us mightily in our next order of business of finding evidence for that very truth. Sherlock Holmes has abductively "inferred" that the stablemaster has stolen the famous racing horse "Silver Blaze" (see the Conan Doyle story with this title) in order to lame it partially. He still has to confirm this conclusion, however, and in that process he is guided by the very content of that abductive conclusion—for instance, in directing his attention to the possibility that the stablemaster had practiced his laming operation on the innocent sheep grazing nearby. He puts a question to the shepherd as to whether anything had been amiss with them of late. "Well, sir, not of much account, but three of them have gone lame, sir." Without having already hit on the truth, Holmes could not have thought of asking this particular question.

If you disregard the strategic angle, the frequent practice of such "jumps to a conclusion" by scientists may easily lead one to believe that scientific discovery is not subject to epistemological rules. The result will then be the hypothetico-deductive model of scientific reasoning, which is hence seen to rest on a fallacious dismissal of the strategic angle.

Thus we reach a result that is neatly contrary to what were once prevalent views. It used to be held that discovery cannot be subject to explicit epistemological theory, whereas justification can. We have found out that not only can discovery be

approached epistemologically, but that justification cannot in the long run be done justice to by a theory that does not also cover discovery.

A critical reader might initially have been wondering why contexts of verification and of other forms of justification do not constitute a third logical home of the notion of knowledge, besides the contexts of decision-making and information-acquisition. The answer is that processes of justification can only be considered as aspects of processes of information-acquisition.

The Generality of the Interrogative Model

The most general argument for the generality of the interrogative approach relies only on the assumption that the inquirer's line of thought can be rationally evaluated. What is needed for such an evaluation? If no new information is introduced into an argument by a certain step, then the outcome of that step is a logical consequence of earlier statements reached in the argument. Hence we are dealing with a logical inference step that has to be evaluated by the criteria of logical validity. It follows that interrogative steps are the ones in which new information enters into the argument. In order to evaluate the step, we must know what the source of this information is, for the reliability of the information may depend on its source. We must also know what else might have resulted from the inquirer's approaching this particular source in this particular way and with what probabilities. If so, what the inquirer did can be thought of as a question addressed to that source of information. Likewise, we must know what other sources of information the inquirer could have consulted and what the different results might have been. This amounts to knowing what other sources of answers the inquirer might have consulted. But if all of this is known, we might as well consider what the inquirer did as a step in interrogative inquiry.

In an earlier work (Hintikka 1998), I have likened such tacit interrogative steps to Peircean abductions, which Peirce insists are inferences even though they have interrogative and conjectural aspects.

The interrogative model can be thought of as having also another kind of generality—namely, generality with respect to the different kinds of questions. Earlier epistemic logic was incapable of handling questions more complicated than simple *wh*-questions. In particular, it could not specify the logical form of questions in which the questioned ingredient was apparently within the scope of a universal quantifier, which in turn was in the scope of a *knows that* operator. This defect was eliminated by means of the independence indicator (slash) /. (See Hintikka 2003b.) What characterizes the questioned ingredient is its independence of the epistemic operator, and such independence is perfectly compatible with its being dependent on a universal quantifier, which is in turn dependent on the universal quantifier. In symbols we can now write, for instance, K $(\forall x)$ $(\exists y/K)$ without having to face the impossible task of capturing the threefold dependence structure by means of scopes—that is, by ordering K, $(\forall x)$, and $(\exists y)$ linearly so as to capture their dependence relations.

In this way, we can treat all *wh*-questions and all prepositional questions (involving questions where the two kinds of question ingredients are intermingled). The question ingredient of prepositional questions turns out to be of the form (v/K) and the question ingredient of *wh*-questions of the form (∃x/K). We can also close a major gap in our argument so far. The connection between knowledge and decision-making discussed in section "Knowledge and decision-making" is apparently subject to the serious objection mentioned in section "The logic of knowledge and information". It helps to understand a knowledge operator K only when it occurs clause-initially, prefixed to a closed sentence. For it is only such sentences, not all and sundry formulas, that express a proposition that can serve as a justification of an action. Occurrences of K inside a sentence prefixed to an open formula cannot be interpreted in the same way. Now we can restrict K to a sentence-initial position, which eliminates this objection. This also helps to fulfill the promise made in section "The logic of knowledge and information" of constructing a general logic for the epistemic operator. Here we are witnessing a major triumph of second-generation epistemic logic, which relies on the notion of independence. It solves once and for all the problem of "quantifying in." It turns out that we do not at bottom *quantify into* a context governed by the epistemic operator K. What we in effect do is to *quantify independently* of this operator.

Why-questions and *how*-questions require a special treatment, which nevertheless is not hard to do. (See, e.g., Hintikka and Halonen 1995.)

The most persuasive argument for the interrogative model nevertheless comes from the applications of the interrogative viewpoint to different problems in epistemology. An important role in such applications is played by the presuppositions of questions and by the presuppositions of answers, better known as their conclusiveness conditions. Examples of such application are offered in Chaps. 4 and 5 of this volume.

The Place of Knowledge in Inquiry

It would take me too far afield here to essay a full-fledged description of the interrogative model. It is nevertheless easy to make an inventory of the concepts that are employed in it. In an explicit model, question-answer steps are interspersed with logical inference steps. Hence the concepts of ordinary deductive logic are needed. As long as the inquirer can trust all the answers, the concepts that are needed are the presuppositions of a question, the conclusiveness condition of an answer (which might be called the "presupposition" of the answer), and the notion of information. To describe an interrogative argument with uncertain answers (responses), we need the notion of tentative rejection of an answer, also known as *bracketing*, and hence also the converse operation of unbracketing, plus ultimately also the notion of probability needed to judge the conditions of bracketing and unbracketing.

What is remarkable about this inventory is that it does not include the concept of knowledge. One can construct a full epistemological theory of inquiry as inquiry

without ever using the k-word. This observation is made especially significant by the generality of the interrogative model. As was indicated, not only is it by means of an interrogative argument that all new information can be thought of as having been discovered, it is by the same questioning method that its credibility must be established in principle.

What this means is that by constructing a theory of interrogative inquiry we apparently can build up a complete theory of epistemology without using the concept of knowledge. We do not need the notion of knowledge in our theory of knowledge—or so it seems. We do not need it either in the theory of discovery or in the theory of justification.

This conclusion might seem to be too strange to be halfway plausible. It is not, but it needs explanations to be seen in the right perspective.

It might perhaps seem that the concept of knowledge is smuggled into interrogative argumentation by the epistemic logic that has to be used in it. This objection is in fact a shrewd one. I said earlier that the logic of questions and answers, which is the backbone of the interrogative model, is part of the logic of knowledge. And this need to resort to epistemic notions is grounded deeply in the facts of the case. It might at first seem that in an interrogative inquiry, no epistemic notions are needed. The presuppositions of questions, questions themselves, and replies to them can apparently be formulated without using epistemic notions.

However, this first impression turns out to be misleading. The structure of and the rules governing it cannot be specified without using some suitable epistemic logic. For one thing, many of the properties of questions and answers are best explained by reference to what is known as the desideratum of a question. This desideratum specifies the epistemic state that the questioner wants to be brought about (in the normal use of questions). For instance, the desideratum of "Who murdered Roger Ackroyd?" is "I know who murdered Roger Ackroyd." But the desideratum with its *prima facie* knowledge operator is not only a part of a theory of question-answer sequences, it is a vital ingredient of the very interrogative process.

In particular, it is needed to solve Meno's problem (Plato 1924) applied to interrogative inquiry. In the initial formulation of the rules for interrogative inquiry, it is apparently required that we must know not only the initial premises of inquiry but also their ultimate conclusion. This seems to mean that we can use interrogative inquiry only to explain conclusions we have already reached but not to solve problems—in other words, answer questions by means of questions. But in trying to answer a question by means of interrogative inquiry, we apparently do not know what the ultimate conclusion is. We are instead looking for it. How, then, can we use interrogative inquiry for the purpose of answering questions? The answer is that we must formulate the logic of inquiry in terms of what the inquirer knows (in the sense of being informed about) at each stage. Then we can solve Meno's problem merely by using the desideratum of the overall question as the ultimate conclusion. But then we seem to need the notion of knowledge with vengeance.

What is true is that a viable theory of questions and answers will inevitably involve an intensional operator, and in particular an epistemic operator in a wide

sense of the word. However, the epistemic attitude this operator expresses is not knowledge in any reasonable sense of the word, not just not in the philosopher's solemn sense. Here, the results reached in section "The logic of knowledge and information" are applicable. Before an interrogative inquiry has reached its aim—that is, knowledge—we are dealing with information that has not yet hardened into knowledge. It was seen earlier that the logic of such unfinished epistemological business is indeed a kind of epistemic logic, but a logic of information rather than of knowledge.

This point is worth elaborating. Indeed the real refutation of the accusation of having smuggled the concept of knowledge into interrogative inquiry in the form of the epistemic operator used in questions and answers lies in pointing out the behavior of this operator in epistemic inquiry. It may sound natural to say that after having received what is known as a conclusive answer to a question, the inquirer now knows it. But the notion of knowledge employed here is a far cry from the notion of knowledge that philosophers have tried to define. It looks much more like the ugly foreign notion of information. It does not even carry the implication of truth, for the answer might very well have to be bracketed later in the same inquiry. By the same token, it does not even presuppose any kind of stable belief in what is "known." Instead of saying that after having received a conclusive answer, the inquirer knows it, it would be more accurate to say that he or she has been informed about it. Here the advantages of the less deep notion of information are amply in evidence. Unlike knowledge, information need not be true. If an item of information offered to me turns out to be false, I can borrow a line from *Casablanca* and ruefully say, "I was misinformed." The epistemic operator needed in the logic of questions and answers is therefore not a knowledge operator in the usual sense of the term. My emphasis on this point is a penance, for I now realize that my statements in the past might have conveyed to my readers a different impression. What is involved in the semantics of questions and answers is the logic of information, not the logic of knowledge. This role of the notion of information in interrogative inquiry is indeed crucial, but it does not involve epistemologists' usual concept of knowledge at all.

This point is so important as to be worth spelling out even more fully. Each answer presents the inquirer with a certain item of information, and the distinction between question-answer steps and logical inferences steps hinges on the question of whether this information must be old or whether it can be new information. But it is important to realize that such information does not amount to knowledge. In an ongoing interrogative inquiry, there are no propositions concerning which question is ever raised, whether they are known or not. There may be a provisional presumption that, barring further evidence, the answers that an inquirer receives are true, but there is not even a whiff of a presumption that they are known. Conversely, when an answer is bracketed, it does not mean that it is definitively declared not to be known, for further answers may lead the inquirer to unbracket it. In sum, it is true in the strictest possible sense that the concept of knowledge in anything like philosophers' sense is not used in the course of interrogative inquiry.

These observations show the place of knowledge in the world of actual inquiry, and it also shows the only context in which questions about the definition of knowledge can legitimately be asked. The notion of knowledge may or may not be a discussion-stopper, but it is certainly an inquiry-stopper.

It might be suspected that this is due to the particular way the interrogative model is set up. Such a suspicion is unfounded, however. The absence of the concept of knowledge from ampliative inquiry is grounded in the very nature of the concept of knowledge. Questions of knowledge do not play any role in the questioning process itself, only in evaluating its results. For what role was it seen to play in human life? It was seen as what justifies us to act in a certain way. The concept of knowledge is therefore related to interrogative inquiry by asking: When has an interrogative inquiry reached far enough to justify the inquirer's acting on the basis of the conclusions it has so far reached? Or, to align this question with the locutions used earlier, when has the inquiry entitled the inquirer to dismiss the scenarios that are incompatible with the propositions accepted in the inquiry at the time? This is a genuine question, and it might seem to bring the concept of knowledge to the center of the theory of interrogative inquiry.

In a sense it does that. But this sense does not bring the notion of knowledge back as a concept that can possibly figure in the definitory rules of inquiry. It brings knowledge back to the sphere of strategic aspects of inquiry. The question as to whether a conclusion of inquiry has been justified strongly enough for it to qualify as knowledge is on a par with the question as to whether or not a step in an inquiry (typically an answer to a question) should perhaps be bracketed (however tentatively). Both are strategic questions. It is hopeless to try to model knowledge acquisition in a way that turns these decisions into questions of definitory correctness.

Any context-free definition of knowledge would amount to a definitory rule in the game of inquiry—namely, a definitory rule for stopping an inquiry. And once one realizes that this is what a definition of knowledge would have to do in the light of the conception of inquiry as inquiry, one realizes that the pursuit of such a definition is a wild goose chase.

It is important to realize that this conclusion does not only apply to attempted definitions of knowledge that refer only to the epistemic situation that has been reached at the putative end stage of the "game" of inquiry. In other words, it does not apply only to the state of an inquirer's evidence at the end of an inquiry. It also applies to definitions in which the entire history of inquiry so far is taken into account.

This conclusion is worth spelling out more fully. What the conclusion says is that no matter how we measure the credence of the output of interrogative inquiry, there is no reason to believe that an answer to the question as to when an inquirer is justified to act on his or her presumed knowledge depends only on the process of inquiry through which the inquirer's information has been obtained independently of the subject matter of the inquiry. In an old terminology, the criteria of justification cannot be purely *ad argumentum*, but must also be *ad hoc*. Neither the amount of

information nor the amount of justification that authorizes an agent to stop his or her inquiry and act on its results can always be specified independently of the subject matter—for instance, independently of the seriousness of the consequences of being wrong about the particular question at hand. And if the justification depends on the subject matter, then so does the concept of knowledge, because of the roots of our concept of knowledge in action.

But since the notion of knowledge was seen to be tied to the justification of acting on the basis of what one knows, the concept of knowledge depends on the subject matter and not only on the epistemological situation. Accordingly, no general definition of knowledge in purely epistemological terms is possible.

This point is not a relativistic one as far as the possibility of *a priori* epistemology is concerned. If anything, the divorce of knowledge from inquiry underlines the objectivity of inquiry and its independence of the value aspects of the subject matter. The fashionable recent emphasis on the alleged value-ladenness of science is misleading in that it is typically predicated on forgetting or overlooking that the question as to when the results of scientific inquiry authorize acting on them is different from questions concerning the methodology of scientific inquiry itself. The dependence of the criteria of knowledge on subject matter ought to be a platitude. It is one thing for Einstein to claim that he knew that the special theory of relativity was true notwithstanding *prima facie* contrary experimental evidence, and another thing for a medical researcher to be in a position to claim to know that a new vaccine is safe enough to be administered to sixty million people. But some relativists mistakenly take this platitude to be a deep truth about scientific methodology and its dependence on subject matter. This is a mistake in the light of the fact that the allegedly value-laden concept of knowledge does not play any role in the actual process of inquiry.

Here, a comparison with such decision principles as the maximization of expected utility is instructive. What an inquiry can provide is only the expectations (probabilities). But they do not alone determine the decision, which depends also on the decider's utilities. Hence the criteria of knowing cannot be defined by any topic-neutral general epistemology alone. But this dependence does not mean that the probabilities used—misleadingly called "subjective" probabilities—should in rational decision-making depend on one's utilities. Decision-making based on such probability estimates would be paradigmatically irrational.

The influence of subject matter on the notion of knowledge does not imply that the interrogative process through which putative knowledge has been obtained is irrelevant for the evaluation of its status. Here lies, in fact, a promising field of work for applied epistemologists. Material for such work is available in, among many other places, different kinds of studies of risk-taking. Even though considerations of strategies do not help us to formulate a topic-neutral definition of knowledge, in such a topic-sensitive epistemology they are bound to play a crucial role. This is a consequence of the general fact that in game-like processes, only strategies, not individual moves, can in the last analysis be evaluated.

Comparisons with Other Epistemologists

Relativizing our humanly relevant concept of knowledge to some particular subject matter also provides a strategy of answering a philosophical skeptic. If knowledge claims depend for their very meaning on the criteria governing some particular walk of human action, then so also must reasonable doubts. It is only unspecific "philosophical" doubts that do not have built into their own logic standards that show how they can be surmounted.

One philosopher who would have agreed with my thesis concerning the dependence of the criteria of knowledge on the subject matter, and who in fact supplied reasons for it, is Ludwig Wittgenstein. In Hintikka forthcoming, I have shown that according to Wittgenstein's mature views, the concept of knowledge cannot be used in what I have called "primary language-games." These language-games are for Wittgenstein the direct links between language and reality. In them, we cannot, in Wittgenstein's metaphor, drive a wedge between language and what it expresses. Such a primary language-game does not operate by means of criteria, but by means of spontaneous responses. If I try to say in such a primary language-game "I know that I am in pain," all that I can express is the same as "I am in pain." And in a primary language-game, to utter "I am in pain" is but a form of pain-behavior.

In Wittgenstein's view, epistemic concepts can be used only in what I have called "secondary language-games." These secondary language-games presuppose primary ones. They do not operate through spontaneous responses, verbal or behavioral, and hence they must involve criteria. For this reason, epistemic vocabulary can be used in them. But those criteria are different in different secondary games. Hence the force of epistemic terms depends on the particular secondary game in which they are being used. Saying this is very nearly nothing but Ludwigspeak for saying that the criteria of knowing depend on the subject matter.

Other epistemologists have not been unaware, either, of connections between the justifiability of knowledge claims and the subject matter involved. (See, e.g., DeRose 1995; Cohen 1998; Williams 2001, ch. 14; Bonjour 2002, pp. 267–271.) They seem to have ascribed the dependence in question to the context of inquiry rather than to its subject matter, however. Unless and until the notion of context used here is clarified, I remain doubtful of such claims of context-dependence. For instance, criteria of knowing that a vaccine is safe depend on the life-or-death character of the subject matter, but they presumably should not depend on the context, which may be an administrative decision to initiate compulsory vaccination or a pharmaceutical company's promise to produce the requisite vaccine. However, if the notion of context is interpreted in such a way that it includes first and foremost the subject matter of inquiry, contextualist epistemology might very well converge with the views expressed here. In this work, contextualism is not examined further, however.

Moreover, contextual epistemologists seem to have assimilated the insight into the context-dependence of knowledge to another insight—namely, to the insight that every epistemological inquiry concerns some particular model, a "system"

as physicists would call it, which typically is not an entire world. (See here Hintikka 2003a.) All epistemological inquiry is therefore contextual in this sense of being relative to a model (scenario or "possible world"). But this does not make epistemology itself contextual or relative as a scientific theory is made contextual or relative by the fact that it is inevitably applied to reality system by system. Hence the impact of the line of thought pursued here is diametrically opposed to the most common form of contextualism. This form of contextualism aims at the rejection of global epistemological questions. (See Bonjour 2002, p. 267). For us, global epistemological questions concern in the first place the nature of interrogative inquiry, and they are in no sense context-dependent or even dependent on the subject matter.

The Folly of Trying to Define Knowledge

The concept of knowledge thus belongs to applied epistemology, not to general epistemology. The criteria of knowledge concern the conditions on which the results of epistemological inquiry can be relied as a basis of action. It follows that it is an exercise in futility to try to define knowledge in any general epistemological theory. Such a definition could never help Satire's sleeper spy. But my point is not only about what is not useful in practice. The extensive discussions about how to define knowledge are not only useless for applications, they are theoretically misguided. Here the true relations of the concepts knowledge and truth to definability are almost precisely opposite to what they have been taken to be recently. Tarski (1956) proved certain results concerning the undefinability of truth. Philosophers and other thinkers have taken Tarski's results at their apparent face value, without realizing how restrictive the assumptions are on which these impossibility results are predicated. (See Hintikka 2002.) They have even let Tarski's results discourage them to the extent of giving up attempts to define truth. Tarski notwithstanding, a truth predicate can be formulated for sufficiently rich languages in a philosophically relevant sense in the same language. In contrast, no major philosopher has to the best of my knowledge openly maintained it to be a folly to try to define knowledge. Yet if someone has done so, that philosopher would have been much closer to truth than a philosopher who argues that it is foolish to try to define truth. (See Davidson 1996.)

Belief as a Product of Inquiry

The notion of knowledge belongs to applied epistemology because it is connected conceptually with the notions of acting and decision-making. The particular connection is not crucial. But if it does not matter, similar conclusions must hold also for those other epistemic concepts that are connected conceptually with behavior,

especially with decision-making. The concept of belief is a case in point. And conclusions similar to the ones that have been reached here concerning the notion of knowledge can in fact be drawn concerning the notion of belief. If you are inspired by this line of thought to review the structure of interrogative inquiry with a view to finding a role for the notion of belief there, you will not find such a role. Receiving an answer and incorporating it into one's interrogative inquiry is not the same as adopting a new belief. Acceptance is not the same thing as belief-formation. (For a discussion of their relationship, see Cohen 1992.) For one thing, at no stage of an interrogative inquiry are there any indications whether or not the inquirer is prepared to act on the truth of the propositions that the inquirer has at that stage accepted (and not bracketed). Hence the entire theory of knowledge acquisition can—and must—be developed without using the notion of belief. This notion does not play any role in an interrogative inquiry, only in the evaluation of its putative end-point. If one thinks about it, the notion of belief does not play much of a role in the methodology of science. What I am suggesting is that is should not play any more of a role in general epistemology either.

There is thus a certain partial epistemological parallelism between belief and knowledge. This parallelism has not been appreciated by epistemologists. Ever since Plato, the two notions are habitually contrasted to each other. This contrast is nevertheless seriously misleading, as far the epistemology of belief is concerned.

It seems to me that the same point is unwittingly attested to by all the decision theorists who are using beliefs as an ingredient in rational decision-making. Such a use would be pointless unless there were some previous reasons to think that the beliefs in question can rationally be expected to be true. And such reasons must somehow come from the inquirer's previous experience, if one is a good empiricist.

Belief, too, is connected with criteria as to when I am ready to act on a certain item of information I have received. But whereas the criteria of knowing are impersonal (even though they can be relative to the subject matter), the criteria of belief can be personal and dependent on an even wider selection of the aspects of the subject matter. In claiming to know, I am making a commitment to others, but in forming a belief, I am usually responsible only to myself.

There are also intermediate cases. For instance, a scientist's beliefs *qua* scientist are subject to the standards of acceptance in his or her scientific community. The crucial point is that those beliefs are, in the case of a scientist, formed as a result of an inquiry, rather than, so to speak, as a response to the question, "What do you think about it?" One may very well catch a physicist asking whether he or she should believe a certain hypothesis in the light of available evidence. But one is even likelier to find a scientific inquirer asking what new information he or she should try to acquire—for instance, what experiments to carry out—in order to be in a position to entertain a certain belief.

In general, the same things can thus be said of belief and its standards as were said earlier of knowledge. Belief statements, like knowledge statements, express entitlement of a certain sort. In neither case does an agent have to avail himself or herself of such entitlement. Beliefs need not manifest themselves in overt behavior any more than knowledge. Hence, decision theorists' frequent assumption that an

agent's beliefs (or degrees of belief) together with utilities determine his, her, or its behavior is in need of scrutiny. Above all, beliefs, too, must be thought of as being formed by means of inquiry.

What I take to be a related point has been expressed by Timothy Williamson by pointing out that a "reason is needed for thinking that beliefs tend to be true." (Quoted from the abstract of his contribution to the conference on "Modalism and Mentalism in Modern Epistemology," Copenhagen, January 29–31, 2004.) The relationship is mediated by the fact that, if I am right, interrogative inquiry is, in the last analysis, the only way of arriving at true beliefs.

The conclusions reached here have repercussions for the entire research strategies that should be pursued in epistemology. For instance, there is a major school of thought that conceives of inquiry as a series of belief revisions. But is this at all realistic as a description of what good reasoners actually do? Georges Simenon's Inspector Maigret is sometimes asked what he believes about the case he is investigating. His typical answer is: "I don't believe anything." And this does not mean, contrary to what one might first suspect, that Maigret wants only to know and not to believe and that he has not yet reached that state of knowledge. No—in one story he says, "The moment for believing or not believing hasn't come yet." (Georges Simenon, *Maigret and the Pickpocket*, Harcourt Brace Jovanovich, San Diego, 1985.) It is not that Maigret has not carried his investigation far enough to be in a position to know something. He has not reached for enough to form a belief. (The mere possibility of using the locution "belief formation" is instructive.) In serious inquiry, belief too is a matter whether an inquiry has reached far enough.

Belief, too, concerns the question of when to stop an inquiry. That is the place of this concept in the framework of the interrogative approach. The difference between belief and knowledge does not lie merely in the degree of justification the believer has reached. It does not mean that there is an evaluative component in knowledge but not in belief. The difference lies in the kind of evaluation involved. It is much more like the difference between satisfying an agent's own freely chosen standards of epistemic confidence and satisfying certain impersonal standards that are appropriate to the subject matter.

In linguists' terminology, knowing is an achievement verb. In a way, although not in a literal sense, believing is in the context of interrogative inquiry likewise an achievement notion. What should be studied in epistemology is belief-formation and not only belief change. The notion of belief cannot serve the role as a determinant of human action that is assigned to it in decision theory if it is not influenced by what the agent knows. But such influence is usually not studied in decision theory.

One corollary to the results we have reached concerns philosophers' research strategies. What we can see now is that the interrogative model is not only a rational reconstruction of knowledge acquisition, it can also be used as a model of belief formation. The insight that belief, too, is typically a product of inquiry lends some renewed interest to the "true belief" type of attempted definitions of knowledge. What they perhaps succeed in capturing is admittedly not philosophers' strong sense of knowledge. But there may be other uses (senses?) of the words *knowledge* and *knowing* that can be approached by means of such characterizations.

Philosophers tend to downplay the role of certainty, especially of experienced certainty, in explicating the notion of knowledge. There is nevertheless a third-person use of knowledge attributions in which the meaning of knowing is very close to true conviction reached through inquiry. In such cases, the inquirer has convinced himself or herself by means of inquiry of the truth of some proposition or other even when, by some standards, the inquirer has not yet reached sufficient justification.

A typical context is when an investigator has reached a correct conclusion—for instance, identified the perpetrator—through inquiry and has become convinced of this conclusion even though his or her reasons would not satisfy the standards of evidence in a court of law. It is interesting to note that in such a usage, the true conclusion must have been reached through a viable strategy. A mere guess would not amount to knowledge even in such cases. (Notice that one could not attribute knowledge in this sense to an automaton or to a database.) This observation may be related to Frank Ramsey's (1978) attempt to characterize knowledge as true belief obtained through a reliable method. This sense of knowing seems to be much closer to colloquial usage than the one philosophers have in vain been trying to define.

Repercussions for Other Approaches

From the point of view we have reached, we can also see some serious problems about the Bayesian approach to inquiry. (See, e.g., Earman 1992.) This approach deals with belief change rather than belief-formation. Insofar as we can find any slot for belief-formation within the Bayesian framework from the point of view of any simple application of, it is pushed back to the selection of priors. In other words, it is made entirely *a priori*, at least locally. This is by itself difficult to implement in the case of theory-formation (belief-formation) in science. Is it, for instance, realistic to assume that a scientist can associate an *a priori* probability with each and every possible law of nature? And these doubts are reinforced by general conceptual considerations. Assignments of priors amount to assumptions concerning the world. What is more, prior probabilities pertain to the entire system (model, "world") that the inquirer is investigating bit by bit. How can the inquirer choose such priors on the basis of his or her limited knowledge of the world? These difficulties might not be crucial if there existed a Bayesian theory of belief-change that included a study of changes of priors. Even though such changes have been studied, it seems to me that their theory has not been developed far enough in the Bayesian framework to cover all possibilities.

All sorts of difficult questions face us here. For instance, in order to use Bayesian inference, we need to know the prior probabilities. It seems to be thought generally that this does not amount to asking very much. This may be true in situations in which the primary data is reasonably reliable, as in typical scientific contexts. However, if our evidence is likely to be relatively unreliable, the situation may be different—for instance, when we are dealing with testimony as our basic form of

evidence. I may easily end up asking: Do I really have enough information to make the guess concerning the world that was seem to be involved in the choice of the priors?

For one thing, even though the matter is highly controversial, fascinating evidence to this effect comes from the theory of so-called cognitive fallacies studied by mathematical psychologists such as Amos Tversky and Daniel Kahneman. (See, e.g., Kahneman et al. 1982; Piatelli-Palmarini 1994.) These alleged fallacies include the conjunction fallacy and the base-rate fallacy. As I have suggested in Chap. 9 of this volume (and in Hintikka 2004), at least in certain "crucial experiment" cases, the alleged mistakes are not fallacious at all, but rather point to certain subtle but very real ways in which one's prior probabilities can (and must) be changed in the light of new evidence. They do not show that certain fallacious ways of thinking are hardwired into human beings. Rather, what they show is that Bayesians have so far failed to master certain subtle modes of ampliative reasoning. Tversky's and Kahneman's Nobel Prize notwithstanding, epistemologists should take a long critical look at the entire theory of cognitive fallacies.

Here I can only give indications of how to view the cognitive fallacies conundrum. Very briefly, in the kind of situation that is at issue in the alleged conjunctive fallacy, the prior probabilities that one in effect relies on include the degrees of probability (credibility) assigned to the reports one receives. But that credibility can not only be affected by suitable new evidence, it can be affected by the very report itself. If the report shows that the reporter is likely to know more about the subject matter than another one, it is not fallacious to assign a higher prior probability to his or her report, even though it is a conjunction of a less credible report and further information.

In the case of an alleged base-rate fallacy, there is no conceivable mistake present if the intended sample space consists simply of the different possible courses of events concerning the crucial event—for example, a traffic accident. Base rates enter into the picture only when a wider class of courses of events is considered—for example, all possible courses of events that might have led to the accident. This means considering a larger sample space. Either sample space can of course be considered entirely consistently, depending on one's purposes. A fallacy would inevitably be committed only if the only legitimate application of our language and our epistemological methods was to the entire world—in this case, the larger sample space. But such an exclusive preference of the larger sample space is but an instance of the one-world assumption, which I have criticized elsewhere. (See Hintikka 2003a.)

Whither Epistemology?

The moral of the insights we have thus reached is not merely to avoid certain words in our epistemological theorizing. It calls for rethinking our overall research strategies in epistemology. And the spirit in which we should do so is perhaps

illustrated by the first epistemologist in the Western philosophical tradition. Socrates did not claim that he knew anything. In the manner of a practitioner of my interrogative method, what he did was to ask questions. I suspect that it is only in Plato's dialogues that he was looking for a definition of knowledge. And Plato put this question (and other questions of definition) into Socrates's mouth because Plato shared the widespread Greek assumption that the definition of X gives us the "blueprint" that enables us to bring about X. (See Hintikka 1974, ch. 1–2.) This applies both to the generic search for knowledge and to the quest of particular items of knowledge. Thus, insofar as Plato contemplated knowledge-seeking (information-seeking) by questioning in our sense, he would have had to say that we must know what we are looking for there and that it is this knowledge alone that can guide our search. (No wonder he was worried about Meno's problem.) By the same token, all search for knowledge would have had to be guided by our knowledge of what knowledge is.

Hence it is seen that Plato had in one important respect the same focus as we: the quest for knowledge rather than the justification of beliefs. The definition of knowledge was thought of by Plato as a means for this quest. If so, the pursuit of the definition of knowledge would indeed have been the alpha and omega of epistemology. But we do not think in that way. The training that Satire's spymaster is supposed to have received did not aim exclusively at learning the definition of knowledge. For us, the fact that knowledge can be considered the end product of inquiry shows on the contrary that it cannot play any role in the process of inquiry. Hence the wild goose chase of the definition of knowledge only shows that too many contemporary epistemologists are still bewitched by Plato's assumptions. This is one of the reasons why at the beginning of this chapter, I called contemporary academic epistemology antiquated. Maybe it is time for its practitioners to take up some more up-to-date problems.

References

Austin, J. (1961a) (original 1946). Other minds. In *Philosophical papers.* Oxford: Clarendon Press, especially pp. 67–68.

Austin, J. (1961b). Performative utterances. In *Philosophical papers.* Oxford: Clarendon Press, ch. 10, especially p. 220.

Bonjour, L. (2002). *Epistemology: Classical problems and contemporary responses*. Lanham: Rowman & Littlefield.

Cohen, L. J. (1992). *Essay on belief and acceptance*. Oxford: Clarendon.

Cohen, S. (1998). Contextual solutions to epistemological problems. *Australasian Journal of Philosophy, 76*, 289–306.

Collingwood, R. G. (1940). *An essay on metaphysics*. Oxford: Clarendon Press, especially ch. I, sec. 5.

Davidson, D. (1996). The folly of trying to define truth. *Journal of Philosophy, 93*, 263–278.

DeRose, K. (1995). Solving the skeptical problem. *Philosophical Review, 104*, 1–52.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge: MIT Press.

Frege, G. (1984) (original 1918–19). Thoughts. In *Collected papers*. Oxford: Basil Blackwell.
Gadamer, H. -G. (1975) (original 1960). *Truth and method.* Continuum, New York, especially the section "Logic of Questions and Answers," pp. 333–341.
Hintikka, J. (1974). *Knowledge and the known*. Dordrecht: Reidel.
Hintikka, J. (1998). What is abduction? The fundamental problem of contemporary epistemology. *Transactions of the Charles Peirce Society*, *34*, 503–533. A revised version, "Abduction—Inference, Conjecture, or an Answer to a Question," appears as Chapter 2 in this volume.
Hintikka, J. (1999). *Inquiry as inquiry: A logic of scientific discovery*. Dordrecht: Kluwer.
Hintikka, J. (2002). Post-tarskian truth. *Synthese, 126*, 17–36.
Hintikka, J. (2003a), A distinction too few or too many: A vindication of the snalytic vs. synthetic distinction. In Carol C. Gould (Ed.), *Constructivism and practice: Toward a historical epistemology* (pp. 47–74). Lanham: Rowman & Littlefield.
Hintikka, J. (2003b). A second-generation epistemic logic and its general significance. In Vincent F. Hendricks, et al. (Eds.), *Knowledge contributors* (pp. 33–56). Dordrecht: Kluwer Academic. And as Chap. 3 in this volume.
Hintikka, J. (2004). A fallacious fallacy? *Synthese, 140*, 25–35. And as Chap. 9 in this volume.
Hintikka, J. (forthcoming). *Wittgenstein on knowledge and skepticism* (Working paper).
Hintikka, J., & Halonen, I. (1995). Semantics and pragmatics for why-questions. *Journal of Philosophy, 92*, 636–657.
Husserl, E. (1983) (original 1913). *Ideas pertaining to a pure phenomenology. First book*. The Hague: Martinus Nijhoff.
Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
Kant, I. (1787). *Kritik der reinen Vernunft*, 2nd ed. (see Preface, p. xiii).
McCarthy, J. (1990). Circumscription—A form of non-monotonic reasoning. *Artificial Intelligence, 13*, 27–39 and 171–172.
Montague, R. (1974). *Formal philosophy*. New Haven: Yale University Press.
Piatelli-Palmarini, M. (1994). *Inevitable illusions*. New York: Wiley.
Plato. (1924). *Meno*, *Plato: With an English translation* (Loeb classical library, vol. IV). Cambridge: Harvard University Press.
Ramsey, F. (1978) (original 1929). Knowledge. In *Foundations: Essays in philosophy, logic, mathematics, and economics* (pp. 126–127). London: Routledge & Kegan Paul.
Safire, W. (1995). *The sleeper spy*. New York: Random House.
Shope, R. K. (1983). *The analysis of knowing: A decade of research*. Princeton: Princeton University Press.
Tarski, A. (1956) (original 1935). The concept of truth in formalized languages. In *Logic, semantics, metamathematics* (pp. 152–278). Oxford: Clarendon Press.
Williams, M. (2001). *Problems of knowledge: A critical introduction to epistemology*. Oxford: Oxford University Press.

Chapter 27
Epistemic Operators

Fred I. Dretske

Suppose Q is a necessary consequence of P. Given only this much, it is, of course, quite trivial that if it is true that P, then it must also be true that Q. If it is a fact that P, then it must also be a fact that Q. If it is necessary that P, then it is necessary that Q; and if it is possible that P, then it must also be possible that Q.

I have just mentioned four prefixes: 'it is true that' 'it is a fact that', 'it is necessary that', and 'it is possible that'. In this paper I shall refer to such affixes as *sentential operators* or simply *operators;* when affixed to a sentence or statement, they operate on it to generate another sentence or statement. The distinctive thing about the four operators I have just mentioned is that, if Q is a necessary consequence of P, then the statement we get by operating on Q with one of these four operators is a necessary consequence of the statement we get by operating on P with the same operator. This may be put more succinctly if we let 'O' stand for the operator in question and 'O(P)' for the statement we get by affixing the operator 'O' to the statement 'P'. We can now say that the above four operators share the following property: if P entails Q, then O(P) entails O(Q). I shall call any operator having this property a *penetrating operator* (Or, when emphasis is required, a fully *penetrating* operator). In operating on P these operators penetrate to every necessary consequence of P.

Versions of this paper were read to the philosophy departments of several universities in the United States and Canada during the year 1969/1970. I profited greatly from these discussions. I wish especially to thank Paul Dietl who helped me to see a number of points more clearly (perhaps still not clearly enough in his opinion). Finally, my exchanges with Mr. Don Affeldt were extremely useful; I am much indebted to him in connection with some of the points made in the latter portions of the paper.

F.I. Dretske (deceased)
University of Wisconsin, Madison, WI, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_27

We are now in a position to ask ourselves a preliminary question. The answer to this question is easy enough, but it will set the stage for more difficult questions. Are all sentential operators fully penetrating operators? Are all operators such that if *P* entails *Q*, then O(*P*) entails O(*Q*)? If *all* operators are penetrating operators, then each of the following statements must be true (when *P* entails *Q*):

1. You cannot have a reason to believe that *P* unless you have a reason to believe that *Q*.
2. You cannot know that *P* unless you know that *Q*.
3. You cannot explain why *P* is the case unless you can explain why *Q* is the case.
4. If you assert that *P*, then you assert that *Q*.
5. If you hope that *P*, then you hope that *Q*.
6. If it is strange (or accidental) that *P*, then it must be strange (or accidental) that Q.
7. If it was a mistake that *P*, then it was a mistake that *Q*.

This list begins with two epistemic operators, I 'reason to believe that' and 'know that'. Since I shall be concerned with these later in the paper, let me skip over them now and look at those appearing near the end of the list. They will suffice to answer our opening question, and their status is much less problematic than that of some of the other operators.

'She lost' entails 'Someone lost'. Yet, it may be strange that she lost, not at all strange that someone lost. 'Bill and Susan married each other' entails that Susan got married; yet, it may be quite odd that (strange that, incredible that) Bill and Susan married each other but quite unremarkable, not at all odd that, Susan got married. It may have been a mistake that they married each other, not a mistake that Susan got married. Or finally, 'I hit the bull's-eye' entails that I either hit the bull's-eye or the side of the barn; and though I admit that it was lucky that (accidental that) I hit the bull's-eye, I will deny that it was lucky, an accident, that I hit either the bull's-eye or the side of the barn.

Such examples show that not all operators are fully penetrating. Indeed, such operators as 'it is strange that', 'it is accidental that' and 'it is a mistake that' fail to penetrate to some of the most elementary logical consequences of a proposition. Consider the entailment between "*P* and *Q*" and '*Q*'. Clearly, it may be strange that *P* and *Q*, not at all strange that *P*, and not at all strange that *Q*. *A* concatenation of factors, no one of which is strange or accidental may itself be strange or accidental. Taken by itself, there is nothing odd or suspicious about Frank's holding a winning ticket in the first race. The same could be said about any of the other races: there is nothing odd or suspicious about Frank's holding a winning ticket in the nth race. Nonetheless, there is something very odd, very suspicious, in Frank's having a winning ticket in n races.

Therefore, not only are these operators not fully penetrating, they lie, as it were, on the other end of the spectrum. They fail to penetrate to some of the most elementary consequences of a proposition. I shall refer to this class of operators as nonpenetrating operators. I do not wish to suggest by this label that such operators are totally impotent in this respect (or that they are all uniform in their degree of

penetration). I mean it, rather, in a rough, comparative, sense: their *degree of penetration* is less than that of any of the other operators I shall have occasion to discuss.

We have, then, two ends of the spectrum with examples from both ends. Anything that falls between these two extremes I shall call a *semi-penetrating* operator. And with this definition I am, finally, in a position to express my main point, the point I wish to defend in the rest of this paper. It is, simply, that all epistemic operators are semi-penetrating operators. There is both a trivial and a significant side to this claim. Let me first deal briefly with the trivial aspect.

The epistemic operators I mean to be speaking about when I say that all epistemic operators are semi-penetrating include the following:

(a) *S* knows that ...
(b) *S* sees (or can see) that ...
(c) *S* has reason (or a reason) to believe that ...
(d) There is evidence to suggest that ...
(e) *S* can prove that ...
(f) *S* learned (discovered, found out) that ...
(g) In relation to our evidence it is probable that ...

Part of what needs to be established in showing that these are all semi-penetrating operators is that they all possess a degree of penetration greater than that of the nonpenetrating operators. This is the trivial side of my thesis. I say it is trivial because it seems to me fairly obvious that if someone knows that *P* and *Q*, has a reason to believe that *P* and *Q*, or can prove that *P* and *Q*, he thereby knows that *Q*, has a reason to believe that *Q*, or can prove (in the appropriate epistemic sense of this term) that *Q*. Similarly, if *S* knows that Bill and Susan married each other, he (must) know that Susan got married (married someone). If he knows that *P* is the case, he knows that *P* or *Q* is the case (where the 'or' is understood in a sense which makes '*P* or *Q*' a necessary consequence of '*P*'). This is not a claim about what it would be appropriate to say, what the person himself thinks he knows or would say he knows. It is a question, simply, of what he knows. It may not be appropriate to *say* to Jim's wife that you know it was either her husband, Jim, or Harold who sent the neighbor lady an expensive gift *when you know it was Harold*. For, although you do know this, it is misleading to say you know it – especially to jim's wife.

Let me accept, therefore, without further argument that the epistemic operators are not, unlike 'lucky that', 'strange that', 'a mistake that' and 'accidental that', nonpenetrating operators. I would like to turn, then, to the more significant side of my thesis. Before I do, however, I must make one point clear lest it convert my entire thesis into something as trivial as the first half of it. When we are dealing with the epistemic operators, it becomes crucial to specify whether the agent in question knows that *P* entails *Q*. That is to say, *P* may entail *Q*, and *S* may know that *P*, but he may not know that *Q because,* and perhaps *only* because, he fails to appreciate the fact that *P* entails *Q*. When *Q* is a simple logical consequence of *P* we do not expect this to happen, but when the propositions become very complex, or the relationship between them very complex, this might easily occur. Let *P* be a set of axioms, *Q* a *theorem*. *S*'s knowing *P* does not entail *S*'s knowing *Q* just because

P entails *Q*; for, of course, *S* may not know that *P* entails *Q*, may not know that *Q* is a theorem. Hence, our epistemic operators will turn out not to be penetrating because, and perhaps only because, the agents in question are not fully cognizant of all the implications of what they know to be the case, can see to be the case, have a reason to believe is the case, and so on. Were we all ideally astute logicians, were we all fully apprised of all the necessary consequences (supposing this to be a well defined class) of every proposition, perhaps then the epistemic operators would turn into fully penetrating operators. That is, assuming that if *P* entails *Q*, we know that *P* entails *Q*, then every epistemic operator is a penetrating operator: the epistemic operators penetrate to all the known consequences of a proposition.

It is this latter, slightly modified, claim that I mean to reject. Therefore, I shall assume throughout the discussion that when *Q* is a necessary consequence of *P*, every relevant agent *knows that it is.* I shall be dealing with only the *known consequences* (in most cases because they are immediate and obvious consequences). What I wish to show is that, even under this special restriction, the epistemic operators are *only* semi-penetrating.

I think many philosophers would disagree with this contention. The conviction is that the epistemic worth of a proposition is hereditary under entailment, that whatever the epistemic worth of *P, at least* the same value must be accorded the known consequences of *P*. This conviction finds expression in a variety of ways. Epistemic logic: if *S* knows that *P*, and knows that *P* entails *Q*, then *S* knows that *Q*. Probability theory: if *A* is probable, and *B* is a logical consequence of *A*, then *B* is probable (relative to the same evidence, of course). Confirmation theory: if evidence *e* tends to confirm hypothesis *h*, then *e* indirectly confirms all the logical consequences of *h*. But perhaps the best evidence in favor of supposing that most philosophers have taken the epistemic operators to be fully penetrating is the way they have argued and the obvious assumptions that structure their arguments. Anyone who has argued in the following way seems to me to be assuming the thesis of penetrability (as I shall call it): if you do not know whether *Q* is true or not, and *P* cannot be true unless *Q* is true, then you (obviously) do not know whether *P* is true or not. *A* slightly more elaborate form of the same argument goes like this: If *S* does not know whether or not *Q* is true, then for all he knows it might be false. If *Q* is false, however, then *P* must also be false. Hence, for all *S* knows, *P* may be false. Therefore, *S* does not know that *P* is true. This pattern of argument is sprinkled throughout the epistemological literature. Almost all skeptical objections trade on it. *S* claims to know that this is a tomato. *A* necessary consequence of its being a tomato is that it is not a clever imitation which only looks and feels (and, if you will, tastes) like a tomato. But *S* does not know that it is not a clever imitation that only looks and feels (and tastes) like a tomato. (I assume here that no one is prepared to argue that anything that looks, feels, and tastes like a tomato to *S must be* a tomato.) Therefore, *S* does not know that this is a tomato. We can, of course, reply with G. E. Moore that we certainly do know it is a tomato (after such an examination) and since tomatoes are not imitations we know that this is not an imitation. It is interesting to note that this reply presupposes the same principle as does the skeptical objection: they both assume that if *S* knows that this is a *P*, and knows that every *P* is a *Q*, then *S* knows

that this is a *Q*. The only difference is that the skeptic performs a modus tollens, Moore a modus ponens. Neither questions the principle itself.

Whether it be a question of dreams or demons, illusions or fakes, the same pattern of argument emerges. If you know this is a chair, you must know that you are not dreaming (or being deceived by a cunning demon), since its being a (real) chair entails that it is not simply a figment of your own imagination. Such arguments assume that the epistemic operators, and in particular the operator 'to know', penetrate to all the known consequences of a proposition. If these operators were not penetrating, many of these objections might be irrelevant. Consider the following exchange:

S: How strange. There are tomatoes growing in my apple tree.
K: That isn't strange at all. Tomatoes, after all, are physical objects and what is so strange about physical objects growing in your apple tree?

What makes K's reply so silly is that he is treating the operator 'strange that' as a fully penetrating operator: it cannot be strange that there are tomatoes growing in the apple tree unless the consequences of this (e.g., there are objects growing in your apple tree) are also strange. Similarly, it may not be at all relevant to object to someone who claims to know that there are tomatoes in the apple tree that he does not know, cannot be absolutely sure, that there are really any material objects. Whether or not this is a relevant objection will depend on whether or not this particular consequence of there being tomatoes in the apple tree is one of the consequences to which the epistemic operators penetrate. What I wish to argue in the remainder of this paper is that the traditional skeptical arguments exploit precisely those consequences of a proposition to which the epistemic operators do not penetrate, precisely those consequences which distinguish the epistemic operators from the fully penetrating operators.

In support of this claim let me begin with some examples which are, I think, fairly intuitive and then turn to some more problematic cases. I shall begin with the operator 'reason to believe that' although what I have to say could be said as well with any of them. This particular operator has the added advantage that if it can be shown to be only semi-penetrating, then many accounts of knowledge, those which interpret it as a form of justified true belief, would also be committed to treating 'knowing that' as a semi-penetrating operator. For, presumably, 'knowing that' would not penetrate any deeper than one's 'reasons for believing that'.

Suppose you have a reason to believe that the church is empty. *Must* you have a reason to believe that it is a church? I am not asking whether you generally have such a reason. I am asking whether one can have a reason to believe the church empty without having a reason to believe that it is a church which is empty. Certainly your reason for believing that the church is empty is not *itself* a reason to believe it is a church; or it *need not* be. Your reason for believing the church to be empty may be that you just made a thorough inspection of it without finding anyone. That is a good reason to believe the church empty. Just as clearly, however, it is not a reason, much less a good reason, to believe that what is empty is a church. The fact is, or so it seems to me, I do not have to have any reason to believe it is a church. Of course,

I would never *say* the church was empty, or that I had a reason to believe that the church was empty, unless I believed, and presumably had a reason for so believing, that *it was* a church which was empty, but this is a presumed condition of my *saying* something, not of my having a reason to believe something. Suppose I had simply assumed (correctly as it turns out) that the building was a church. Would this show that I had no reason to believe that the church was empty?

Suppose I am describing to you the "adventures" of my brother Harold. Harold is visiting New York for the first time, and he decides to take a bus tour. He boards a crowded bus and immediately takes the last remaining scat. The little old lady he shouldered aside in reaching his seat stands over him glowering. Minutes pass. Finally, realizing that my brother is not going to move, she sighs and moves resignedly to the back of the bus. Not much of an adventure, but enough, I hope, to make my point. I said that the little old lady realized that my brother would not move. Does this imply that she realized that, or knew that, *it was my brother* who refused to move? Clearly not. We can say that *S* knows that *X is Y* without implying that *S* knows that *it is X* which is *Y*. We do not *have* to describe our little old lady as knowing that *the* man or *the person* would not move. We can say that she realized that, or knew that, my brother would not move (minus, of course, this pattern of emphasis), and we can say this because saying this does not entail that the little old lady knew that, or realized that, it was my brother who refused to move. She knew that my brother would not move, and she knew this despite the fact that she did not know something that was necessarily implied by what she did know–viz., that the person who refused to move was my brother.

I have argued elsewhere that to see that *A is B*, that the roses are wilted for example, is not to see, not even to be able to see, that they are roses which are wilted.[1] To see that the widow is limping is not to see that it is a widow who is limping. I am now arguing that this same feature holds for all epistemic operators. I can know that the roses are wilting without knowing that they are roses, know that the water is boiling without knowing that it is water, and prove that the square root of 2 is smaller than the square root of 3 and, yet, be unable to prove what is entailed by this–viz., that the number 2 *has* a square root.

The general point may be put this way: there are certain presuppositions associated with a statement. These presuppositions, although their truth is entailed by the truth of the statement, are not part of what is *operated* on when we operate on the statement with one of our epistemic operators. The epistemic operators do not *penetrate to* these presuppositions. For example, in saying that the coffee is boiling I assert that the coffee is boiling, but in asserting this I do not assert that *it is* coffee which is boiling. Rather, this is taken for granted, assumed, presupposed, or what have you. Hence, when I say that I have a reason to believe that the coffee is boiling, I am not saying that this reason applies to the fact that it is coffee which is boiling. This is *still* presupposed. I may have such a reason, of course, and chances are good

[1] *Seeing and Knowing* (Chicago: University Press, 1969), pp. 93–112, and also "Reasons and Consequences," *Analysis* (April 1968).

that I do have such a reason or I would not have referred to what I believe to be boiling as *coffee,* but to have a reason to believe the coffee is boiling is not, thereby, to have a reason to believe it is coffee which is boiling.

One would expect that if this is true of the semi-penetrating operators, then it should also be true of the nonpenetrating operators. They also should fail to reach the presuppositions. This is exactly what we find. It may be accidental that the two trucks collided, but not at all accidental that it was two trucks that collided. Trucks were the only vehicles allowed on the road that day, and so it was not at all accidental or a matter of chance that the accident took place between two trucks. Still, it was an accident that the two trucks collided. Or suppose Mrs. Murphy mistakenly gives her cat some dog food. It need not be a mistake that she gave the food to *her* cat, or *some* food to a cat. This was intentional. What was a mistake was that it was dog food that she gave to her cat.

Hence, the first class of consequences that differentiate the epistemic operators from the fully penetrating operators is the class of consequences associated with the presuppositions of a proposition. The fact that the epistemic operators do not penetrate to these presuppositions is what helps to make them semi-penetrating. And this is an extremely important fact. For it would appear that if this is true, then to know that the flowers are wilted I do not have to know that they are flowers (which are wilted) and, therefore, do not have to know all those consequences which follow from the fact that they are flowers, real flowers, which I know to be wilted.

Rather than pursue this line, however, I would like to turn to what I consider to be a more significant set of consequences– "more significant" because they are the consequences that are directly involved in most skeptical arguments. Suppose we assert that *x* is *A*. Consider some predicate, '*B*', which is incompatible with *A*, such that nothing can be both *A* and *B*. It then follows from the fact that *x is A* that *x* is not *B*. Furthermore, if we conjoin *B* with any other predicate, *Q*, it follows from the fact that *x is A* that *x* is not(*B* and *Q*). I shall call this type of consequence a *contrast consequence,* and I am interested in a particular subset of these; for I believe the most telling skeptical objections to our ordinary knowledge claims exploit a particular set of these contrast consequences. The exploitation proceeds as follows: someone purports to know that *x is A,* that the wall is red, say. The skeptic now finds a predicate '*B*' that is incompatible with '*A*'. In this particular example we may let '*B*' stand for the predicate 'is white'. Since '*x* is red' entails '*x* is not white' it also entails that *x* is not-(white and *Q*) where '*Q*' is any predicate we care to select. Therefore, the skeptic selects a '*Q*' that gives expression to a condition or circumstance under which a white wall would appear exactly the same as a red wall. For simplicity we may let '*Q*' stand for: 'cleverly illuminated to look red'. We now have this chain of implications: '*x* is red' entails '*x* is not white' entails '*x* is not white cleverly illuminated to look red'. If 'knowing that' is a penetrating operator, then if anyone knows that the wall is red he must know that it is not white cleverly illuminated to look red. (I assume here that the relevant parties know that if *x* is red, it cannot be white made to look red.) He must know that this particular contrast consequence is true. The question is: do we, generally speaking, know anything of the sort? Normally we never take the trouble to check the lighting.

We seldom acquire any *special* reasons for believing the lighting normal although we can talk vaguely about there being no reason to think it unusual. The fact is that we habitually take such matters for granted, and although we normally have good reasons for making such routine assumptions, I do not think these reasons are sufficiently good, not without special precautionary checks in the particular case, to say of the particular situation we are in that we know conditions are normal. To illustrate, let me give you another example–a silly one, but no more silly than a great number of skeptical arguments with which we are all familiar. You take your son to the zoo, see several zebras, and, when questioned by your son, tell him they are zebras. Do you know they are zebras? Well, most of us would have little hesitation in saying that we did know this. We know what zebras look like, and, besides, this is the city zoo and the animals are in a pen clearly marked "Zebras." Yet, something's being a zebra implies that it is not a mule and, in particular, not a mule cleverly disguised by the zoo authorities to look like a zebra. Do you know that these animals are not mules cleverly disguised by the zoo authorities to look like zebras? If you are tempted to say "Yes" to this question, think a moment about what reasons you have, what evidence you can produce in favor of this claim. The evidence you *had* for thinking them zebras has been effectively neutralized, since it does not count toward their not being mules cleverly disguised to look like zebras. Have you checked with the zoo authorities? Did you examine the animals closely enough to detect such a fraud? You might do this, of course, but in most cases you do nothing of the kind. You have some general uniformities on which you rely, regularities to which you give expression by such remarks as, "That isn't very likely" or "Why should the zoo authorities do that?" Granted, the hypothesis (if we may call it that) is not very plausible, given what we know about people and zoos. But the question here is not whether this alternative is plausible, not whether it is more or less plausible than that there are real zebras in the pen, but whether *you know* that this alternative hypothesis is false. I don't think you do. In this I agree with the skeptic. I part company with the skeptic only when he concludes from this that, therefore, you do not know that the animals in the pen are zebras. I part with him because I reject the principle he uses in reaching this conclusion–the principle that if you do not know that Q is true, when it is known that P entails Q, then you do not know that P is true.

What I am suggesting is that we simply admit that we do not know that some of these contrasting "skeptical alternatives" are not the case, but refuse to admit that we do not know what we originally said we knew. My knowing that the wall is red certainly entails that the wall is red; it also entails that the wall is not white and, in particular, it entails that the wall is not white cleverly illuminated to look red. But it does not follow from the fact that I know that the wall is red that I *know* that it is not white cleverly illuminated to look red. Nor does it follow from the fact that I know that those animals are zebras that I know that they are not mules cleverly disguised to look like zebras. These are some of the contrast consequences to which the epistemic operators do not penetrate.

Aside from asserting this, what arguments can be produced to support it? I could proceed by multiplying examples, but I do not think that examples alone will support the full weight of this view. The thesis itself is sufficiently counterintuitive

to render controversial most of the crucial examples. Anyone who is already convinced that skepticism is wrong and who is yet troubled by the sorts of skeptical arguments I have mentioned will, no doubt, take this itself as an argument in favor of my claim that the epistemic operators are only semi-penetrating. This, however, hardly constitutes an argument against skepticism. For this we need *independent* grounds for thinking that the epistemic operators do not penetrate to the contrast consequences. So I shall proceed in a more systematic manner. I shall offer an analogy with three other operators and conclude by making some general remarks about what I think can be learned from this analogy. The first operator is 'explains why' or, more suggestively (for the purposes of this analogy):

(A) *R* is the reason (explanatory reason) that (or why) ...

For example, the reason why *S* quit smoking was that he was afraid of getting cancer. The second operator has to do with reasons again, but in this case it is a reason which tends to *justify* one in doing something:

(B) *R* is a reason for ... (*S* to do *Y*).[2]

For example, the fact that they are selling the very same (type of) car here much more cheaply than elsewhere is a reason to buy it here rather than elsewhere. The status of this as a reason will, of course, depend on a variety of circumstances, but situations can easily be imagined in which this would be a reason for someone to buy the car here. Finally, there is a particular modal relationship which may be construed as a sentential operator:

(C) *R* would not be the case unless ...

For example, he would not have bid seven no-trump unless he had all four aces. I shall abbreviate this operator as '$R\Rightarrow \ldots$'; hence, our example could be written 'he bid seven no-trump $\Rightarrow$ he had all four aces'.

Each of these operators has features similar to those of our epistemic operators. If one retraces the ground we have already covered, one will find, I think, that these operators all penetrate deeper than the typical nonpenetrating operator. If *R* explains why (or is the reason that) *P* and *Q* are the case, then it explains why (is the reason that) *Q* is the case.[3] If I can explain why Bill and Harold are always invited to every party, I can explain why Harold is always invited to every party. From the fact that it was a mistake for me to quit my job it does not follow that it was a mistake for

[2]Unlike our other operators, this one does not have a propositional operand. Despite the rather obvious differences between this case and the others, I still think it useful to call attention to its analogous features.

[3]One must be careful not to confuse sentential conjunction with similar sounding expressions involving a relationship between two things. For example, to say Bill and Susan got married (if it is intended to mean that they married *each other),* although it entails that Susan got married, does not do so by sim*plification.* 'Reason why' penetrates through logical simplification, not through the type of entailment represented by these two propositions. That is, the reason they got married is that they loved each other; that they loved each other is not the reason Susan got married.

me to do something, but if I had a reason to quit my job, it does follow that I had a reason to do something. And if the grass would not be green unless it had plenty of sunshine and water, it follows that it would not be green unless it had water.

Furthermore, the similarities persist when one considers the presuppositional consequences. I argued that the epistemic operators fail to penetrate to the presuppositions; the above three operators display the same feature. In explaining why he takes his lunch to work, I do not (or need not) explain why he goes to work or why he works at all. The explanation may be obvious in some cases, of course, but the fact is I need not be able to explain why he works (he is so wealthy) to explain why he takes his lunch to work (the cafeteria food is *so* bad). The reason why the elms on Main Street are dying is not the reason there are elms on Main Street. I have a reason to feed my cat, no reason (not, at least, the same reason) to have a cat. And although it is quite true that he would not have known about our plans if the secretary had not told him, it does not follow that he would not have known about our plans if *someone other than the secretary* had told him. That is, (He knew about our plans) $\Rightarrow$ (The secretary told him) even though it is not true that (He knew about our plans) $\Rightarrow$ (It was the secretary who told him). Yet, the fact that *it was the secretary* who told him is (I take it) a presuppositional consequence of the fact that *the secretary told* him. Similarly, if George is out to set fire to the first empty building he finds, it may be true to say that George would not have set fire to the church unless it (the church) was empty, yet false to say that George would not have set fire to the church unless *it was a church.*

I now wish to argue that these three operators do not penetrate to a certain set of contrast consequences. To the extent that the epistemic operators are similar to these operators, we may then infer, by analogy, that they also fail to penetrate to certain contrast consequences. This is, admittedly, a weak form of argument, depending as it does on the grounds there are for thinking that the above three operators and the epistemic operators share the same logic in this respect. Nonetheless, the analogy is revealing. Some may even find it persuasive.[4]

(A) The pink walls in my living room clash with my old green couch. Recognizing this, I proceed to paint the walls a compatible shade of green. This is the reason I have, and give, for painting the walls green. Now, in having this explanation for why I painted the walls green, I do not think I have an explanation for two other things, both of which are entailed by what I do have an explanation for. I have not explained why I did not, *instead* of painting the walls green, buy a new couch or cover the old one with a suitable slipcover. Nor have I explained why,

[4]I think that those who are inclined to give a causal account of knowledge should be particularly interested in the operator '$R \Rightarrow \ldots$' since, presumably, it will be involved in many instances of knowledge ("many" not "all," since one might wish to except some form of immediate knowledge-knowledge of one's own psychological state-from the causal account). If this operator is only semipenetrating, then any account of knowledge that relies on the relationship expressed by this operator (as I believe causal accounts must) will be very close to giving a "semi-penetrating" account of 'knowing that'.

instead of painting the walls green, I did not paint them white and illuminate them with green light. The same effect would have been achieved, the same purpose would have been served, albeit at much greater expense.

I expect someone to object as follows: although the explanation given for painting the walls green does not, by itself, explain why the couch was not changed instead, it nonetheless succeeds as an explanation for why the walls were painted green only in so far as there is an explanation for why the couch was not changed instead. If there is no explanation for why I did not change the couch instead, there has been no real, no complete, examination for why the walls were painted green.

I think this objection wrong. I may, of course, have an explanation for why I did not buy a new couch: I love the old one or it has sentimental value. But then again I may not. It just never occurred to me to change the couch; or (if someone thinks that its not occurring to me *is* an explanation of why I did not change the couch) I may have thought of it but decided, for what reasons (if any) I cannot remember, to keep the couch and paint the walls. That is to say, I cannot explain why I did not change the couch. I thought of it but I did not do it. I do not know why. Still, I *can* tell you why I painted the walls green. They clashed with the couch.

(B) The fact that they are selling Xs so much more cheaply here than elsewhere may be a reason to buy your Xs here, but it certainly need not be a reason to do what is a necessary consequence of buying your Xs here–viz., not *stealing your* Xs here.

(C) Let us suppose that S is operating in perfectly normal circumstances, a set of circumstances in which it is true to say that the wall he sees would not (now) look green to him unless it was green (if it were any other color it would look different to him). Although we can easily imagine situations in which this is true, it does not follow that the wall would not (now) look green to S if it were white cleverly illuminated to look green. That is,

(i) The wall looks green (to S) $\Rightarrow$ the wall is green.
(ii) The wall is green *entails* the wall is not white cleverly illuminated to look green (to S).

are both true; yet, it is *not true* that

(iii) The wall looks green (to S) $\Rightarrow$ the wall is not white cleverly illuminated to look green (to S).

There are dozens of examples that illustrate the relative impenetrability of this operator. We can truly say that A and B would not have collided if B had not swerved at the last moment and yet concede that they would have collided without any swerve on the part of B *if* the direction in which A was moving had been suitably altered in the beginning.[5]

[5]The explanation for why the modal relationship between R and P ($R \Rightarrow P$) fails to carry over (penetrate) to the logical consequences of P *(i.e.,* $R \Rightarrow Q$ where Q is a logical consequence of P)

The structure of these cases is virtually identical with that which appeared in the case of the epistemic operators, and I think by looking just a little more closely at this structure we can learn something very fundamental about our class of epistemic operators and, in particular, about what it means to know something. If I may put it this way, within the context of these operators no fact is an island. If we are simply rehearsing the facts, then we can say that it is a fact that Brenda did not take any dessert (though it was included in the meal). We can say this without a thought about what sort of person Brenda is or what she might have done had she ordered dessert. However, if we put this fact into, say, an explanatory context, if we try to explain this fact, it suddenly appears within a network of related facts, a network of possible alternatives which serve to define *what it is that is being explained.* What is being explained is a function of two things-not only the fact (Brenda did not order any dessert), but also the range of relevant alternatives. *A* relevant alternative is an alternative that might have been realized in the existing circumstances if the actual state of affairs had not materialized.[6] When I explain why Brenda did not order any dessert by saying that she was full (was on a diet, did not like anything on the dessert menu), I explain why she did not order any dessert rather *than, as opposed to, or instead of* ordering some dessert and eating *it.* It is this competing possibility which helps to define what it is that I am explaining when I explain why Brenda did not order any dessert. Change this contrast, introduce a different set of relevant alternatives, and you change what it is that is being explained and, therefore, what counts as an explanation, even though (as it were) the same fact is being explained. Consider the following contrasts: ordering some dessert and throwing it at the waiter; ordering some dessert and taking it home to a sick friend.

---

is to be found in the set of circumstances that are taken as *given, or held fixed,* in subjunctive conditionals. There are certain logical consequences of *P* which, by bringing in a reference to circumstances tacitly held fixed in the original subjunctive ($R \Rightarrow P$), introduce a possible variation in these circumstances and, hence, lead to a *different* framework of fixed conditions under which to assess the truth of R @ Q. For instance, in the last example in the text, when it is said that *A* and *B* would not have collided if *B* had not swerved at the last moment, the truth of this conditional clearly takes it *as given that A* and *B* possessed the prior trajectories they in fact had on the occasion in question. Given certain facts, including the fact that they were traveling in the direction they were, they would not have collided if *B* had not swerved. Some of the logical consequences of the statement that *B* swerved do not, however, leave these conditions unaltered–e.g., *B* did not move in a perfectly straight line in a direction $2'$ counterclockwise to the direction it actually moved. This consequence "tinkers" with the circumstances originally taken *as given* (held fixed), and a failure of penetration will usually arise when this occurs. It *need not be* true that *A* and *B* would not have collided if *B* had moved in a perfectly straight line in a direction $2'$ counterclockwise to the direction it actually moved.

[6]I am aware that this characterization of "a relevant alternative" is not, as it stands, very illuminating. I am not sure I can make it more precise. What I am after can be expressed this way: if Brenda had ordered dessert, she *would not* have thrown it at the waiter, stuffed it in her shoes, or taken it home to a sick friend (she has no sick friend). These are not alternatives that might have been realized in the existing circumstances if the actual state of affairs had not materialized. Hence, they are not relevant alternatives. In other words, the 'might have been' in my characterization of a relevant alternative will have to be unpacked in terms of counterfactuals.

With these contrasts none of the above explanations are any longer explanations of why Brenda did not order dessert. Anyone who really wants to know why Brenda did not order dessert and throw it at the waiter will not be helped by being told that she was full or on a diet. This is only to say that, within the context of explanation and within the context of our other operators, the proposition on which we operate must be understood as embedded within a matrix of relevant alternatives. We explain why *P*, but we do so within a framework of competing alternatives *A, B,* and C. Moreover, if the possibility D is not within this contrasting set, not within this network of relevant alternatives, then even though not-D follows necessarily from the fact, *P*, which we do explain, we do not explain why not-D. Though the fact that Brenda did not order dessert and throw it at the waiter follows necessarily from the fact that she did not order dessert (the fact that is explained), this necessary consequence is not explained by the explanation given. The only contrast consequences to which this operator penetrates are those which figured in the original explanation as relevant alternatives.

So it is with our epistemic operators. To know that *x* is *A* is to know that *x is A* within a framework of relevant alternatives, *B, C,* and D. This set of contrasts, together with the fact that *x is A*, serve to define what it is that is known when one knows that *x is A*. One cannot change this set of contrasts without changing what a person is said to know when he is said to know that *x is A*. We have subtle ways of shifting these contrasts and, hence, changing what a person is said to know *without changing the sentence that we use to express what he knows.* Take the fact that Lefty killed Otto. By changing the emphasis pattern we can invoke a different set of contrasts and, hence, alter what it is that *S* is said to know when he is said to know that Lefty killed Otto. We can say, for instance, that *S* knows that *Lefty* killed Otto. In this case (and I think this is the way we usually hear the sentence when there is no *special* emphasis) we are being told that *S* knows the identity of Otto's killer, that *it was Lefty* who killed Otto. Hence, we expect *S*'s reasons for believing that Lefty killed Otto to consist in facts that single out Lefty as the assailant *rather than* George, Mike, or someone else. On the other hand, we can say that *S* knows that Lefty *killed* Otto. In this case we are being told that *S* knows *what Lefty did to Otto;* he killed him *rather than* merely injuring him, killed him *rather than* merely threatening him, etc. *A* good reason for believing that Lefty *killed* Otto (rather than merely injuring him) is that Otto is dead, but this is not much of a reason, if it is a reason at all, for believing that *Lefty* killed Otto. Changing the set of contrasts (from 'Lefty rather than George or Mike' to 'killed rather than injured or threatened') by shifting the emphasis pattern changes what it is that one is alleged to know when one is said to know that Lefty killed Otto.[7] The same point can be made here as we made in the case of explanation: the operator will penetrate only to those contrast consequences which form part of the network of relevant alternatives structuring

[7]The same example works nicely with the operator '$R\Rightarrow \ldots$'. **It** may be true to say that Otto would not be dead unless Lefty killed him (unless what Lefty did **to** him was kill him) without its being true that **Otto** would not be dead unless *Lefty* killed him (unless it was Lefty who killed him).

the original context in which a knowledge claim was advanced. just as we have not explained why Brenda did not order some dessert and throw it at the waiter when we explained why she did not order some dessert (although what we have explained-her not ordering any dessert-entails this), so also in knowing that Lefty *killed* Otto (knowing that what Lefty did to Otto was kill him) we do not nec*essarily* (although we may) know that *Lefty* killed Otto (know that *it was Lefty* who killed Otto). Recall the example of the little old lady who knew that my brother would not move without knowing that it was my brother who would not move.

The conclusions to be drawn are the same as those in the case of explanation. Just as we can say that within the original setting, within the original framework of alternatives that defined what we were trying to explain, we *did explain* why Brenda did not order any dessert, so also within the original setting, within the set of contrasts that defined what it was we were claiming to know, we did know that the wall was red and *did know* that it was a zebra in the pen.

To introduce a novel and enlarged set of alternatives, as the skeptic is inclined to do with our epistemic claims, is to exhibit consequences of what we know, or have reason to believe, which we may not know, may not have a reason to believe; but it does not show that we did not know, did not have a reason to believe, whatever it is that has these consequences. To argue in this way is, I submit, as much a mistake as arguing that we have not explained why Brenda did not order dessert (within the original, normal, setting) because we did not explain why she did not order some and throw it at the waiter.

Chapter 28
Elusive Knowledge

David Lewis

We know a lot. I know what food penguins eat. I know that phones used to ring, but nowadays squeal, when someone calls up. I know that Essendon won the 1993 Grand Final. I know that here is a hand, and here is another.

We have all sorts of everyday knowledge, and we have it in abundance. To doubt that would be absurd. At any rate, to doubt it in any serious and lasting way would be absurd; and even philosophical and temporary doubt, under the influence of argument, is more than a little peculiar. It is a Moorean fact that we know a lot. It is one of those things that we know better than we know the premises of any philosophical argument to the contrary.

Besides knowing a lot that is everyday and trite, I myself think that we know a lot that is interesting and esoteric and controversial. We know a lot about things unseen: tiny particles and pervasive fields, not to mention one another's underwear. Sometimes we even know what an author meant by his writings. But on these questions, let us agree to disagree peacefully with the champions of "post-knowledgeism." The most trite and ordinary parts of our knowledge will be problem enough.

For no sooner do we engage in epistemology – the systematic philosophical examination of knowledge – than we meet a compelling argument that we know next to nothing. The sceptical argument is nothing new or fancy. It is just this: it seems as if knowledge must be by definition infallible. If you claim that S knows that P, and yet you grant that S cannot eliminate a certain possibility in which not-P, it certainly seems as if you have granted that S does not after all know that P.

David Lewis was deceased at the time of publication.

D. Lewis (deceased)
Princeton University, Princeton, NJ, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_28

To speak of fallible knowledge, of knowledge despite uneliminated possibilities of error, just sounds contradictory.

Blind Freddy can see where this will lead. Let your paranoid fantasies rip – CIA plots, hallucinogens in the tap water, conspiracies to deceive, old Nick himself – and soon you find that uneliminated possibilities of error are everywhere. Those possibilities of error are farfetched, of course, but possibilities all the same. They bite into even our most everyday knowledge. We never have infallible knowledge.

Never – well, hardly ever. Some say we have infallible knowledge of a few simple, axiomatic necessary truths; and of our own present experience. They say that I simply cannot be wrong that a part of a part of something is itself a part of that thing; or that it seems to me now (as I sit here at the keyboard) exactly as if I am hearing clicking noises on top of a steady whirring. Some say so. Others deny it. No matter; let it be granted, at least for the sake of the argument. It is not nearly enough. If we have only that much infallible knowledge, yet knowledge is by definition infallible, then we have very little knowledge indeed – not the abundant everyday knowledge we thought we had. That is still absurd.

So we know a lot; knowledge must be infallible; yet we have fallible knowledge or none (or next to none). We are caught between the rock of fallibilism and the whirlpool of scepticism. Both are mad!

Yet fallibilism is the less intrusive madness. It demands less frequent corrections of what we want to say. So, if forced to choose, I choose fallibilism. (And so say all of us.) We can get used to it, and some of us have done. No joy there – we know that people can get used to the most crazy philosophical sayings imaginable. If you are a contented fallibilist, I implore you to be honest, be naive, hear it afresh. "He knows, yet he has not eliminated all possibilities of error." Even if you've numbed your ears, doesn't this overt, explicit fallibilism still sound wrong?

Better fallibilism than scepticism; but it would be better still to dodge the choice. I think we can. We will be alarmingly close to the rock, and also alarmingly close to the whirlpool, but if we steer with care, we can – just barely – escape them both. Maybe epistemology is the culprit. Maybe this extraordinary pastime robs us of our knowledge. Maybe we do know a lot in daily life; but maybe when we look hard at our knowledge, it goes away. But only when we look at it harder than the sane ever do in daily life; only when we let our paranoid fantasies rip. That is when we are forced to admit that there always are uneliminated possibilities of error, so that we have fallible knowledge or none. Much that we say is context-dependent, in simple ways or subtle ways. Simple: "it's evening" is truly said when, and only when, it is said in the evening. Subtle: it could well be true, and not just by luck, that Essendon played rottenly, the Easybeats played brilliantly, yet Essendon won. Different contexts evoke different standards of evaluation. Talking about the Easybeats we apply lax standards, else we could scarcely distinguish their better days from their worse ones. In talking about Essendon, no such laxity is required. Essendon won because play that is rotten by demanding standards suffices to beat play that is brilliant by lax standards.

Maybe ascriptions of knowledge are subtly context-dependent, and maybe epistemology is a context that makes them go false. Then epistemology would be

an investigation that destroys its own subject matter. If so, the sceptical argument might be flawless, when we engage in epistemology-and only then![1]

If you start from the ancient idea that justification is the mark that distinguishes knowledge from mere opinion (even true opinion), then you well might conclude that ascriptions of knowledge are context-dependent because standards for adequate justification are context-dependent. As follows: opinion, even if true, deserves the name of knowledge only if it is adequately supported by reasons; to deserve that name in the especially demanding context of epistemology, the arguments from supporting reasons must be especially watertight; but the special standards of justification that this special context demands never can be met (well, hardly ever). In the strict context of epistemology we know nothing, yet in laxer contexts we know a lot.

But I myself cannot subscribe to this account of the context-dependence of knowledge, because I question its starting point. I don't agree that the mark of knowledge is justification.[2] First, because justification is not sufficient: your true opinion that you will lose the lottery isn't knowledge, whatever the odds. Suppose you know that it is a fair lottery with one winning ticket and many losing tickets, and you know how many losing tickets there are. The greater the number of losing tickets, the better is your justification for believing you will lose. Yet there is no number great enough to transform your fallible opinion into knowledge – after all, you just might win. No justification is good enough – or none short of a watertight deductive argument, and all but the sceptics will agree that this is too much to demand.[3]

Second, because justification is not always necessary. What (non-circular) argument supports our reliance on perception, on memory, and on testimony?[4] And yet we do gain knowledge by these means. And sometimes, far from having supporting arguments, we don't even know how we know. We once had evidence, drew conclusions, and thereby gained knowledge; now we have forgotten our

[1]The suggestion that ascriptions of knowledge go false in the context of epistemology is to be found in Barry Stroud, "Understanding Human Knowledge in General" in Marjorie Clay and Keith Lehrer (eds.), Knowledge and Skepticism (Boulder: Westview Press, 1989); and in Stephen Hetherington, "Lacking Knowledge and lustification by Theorising About Them" (lecture at the University of New South Wales, August 1992). Neither of them tells the story just as I do, however it may be that their versions do not conflict with mine.

[2]Unless, like some, we simply define "justification" as "whatever it takes to turn true opinion into knowledge" regardless of whether what it takes turns out to involve argument from supporting reasons.

[3]The problem of the lottery was introduced in Henry Kyburg, Probability and the Logic of Rational Belief (Middletown, CT: Wesleyan University Press, 1961), and in Carl Hempel, "Deductive-Nomological vs. Statistical Explanation" in Herbert Feigl and Grover Maxwell (eds.), Minnesota Studies in the Philosophy of Science, Vol. II (Minneapolis: University of Minnesota Press, (962). It has been much discussed since, as a problem both about knowledge and about our everyday, non-quantitative concept of belief.

[4]The case of testimony is less discussed than the others; but see C. A. J. Coady, Testimony: A Philosophical Study (Oxford: Clarendon Press, 1992) pp. 79–129.

reasons, yet still we retain our knowledge. Or we know the name that goes with the face, or the sex of the chicken, by relying on subtle visual cues, without knowing what those cues may be.

The link between knowledge and justification must be broken. But if we break that link, then it is not – or not entirely, or not exactly – by raising the standards of justification that epistemology destroys knowledge. I need some different story.

To that end, I propose to take the infallibility of knowledge as my starting point.[5] Must infallibilist epistemology end in scepticism? Not quite. Wait and see. Anyway, here is the definition. Subject S knows proposition P iff (that is, if and only if) P holds in every possibility left uneliminated by S's evidence; equivalently, iff S's evidence eliminates every possibility in which not-P.

The definition is short, the commentary upon it is longer. In the first place, there is the proposition, P. What I choose to call "propositions" are individuated coarsely, by necessary equivalence. For instance, here is only one necessary proposition. It holds in every possibility; hence in every possibility left uneliminated by S's evidence, no matter who S may be and no matter what his evidence may be. So the necessary proposition is known always and everywhere. Yet this known proposition may go unrecognised when presented in impenetrable linguistic disguise, say as the proposition that every even number is the sum of two primes. Likewise, the known proposition that I have two hands may go unrecognised when presented as the proposition that the number of my ands is the least number n such that every even number is the sum of n primes. (Or if you doubt the necessary existence of numbers, switch to an example involving equivalence by logic alone.) These problems of disguise shall not concern us here. Our topic is modal, not hyperintensional, epistemology.[6]

Next, there are the possibilities. We needn't enter here into the question whether these are concreta, abstract constructions, or abstract simples. Further, we needn't decide whether they must always be maximally specific possibilities, or whether they need only be specific enough for the purpose at hand. A possibility will be specific enough if it cannot be split into subcases in such a way that anything we have said about possibilities, or anything we are going to say before we are done, applies to some subcases and not to others. For instance, it should never happen that proposition P holds in some but not all subcases; or that some but not all sub-cases are eliminated by S's evidence.

But we do need to stipulate that they are not just possibilities as to how the whole world is; they also include possibilities as to which part of the world is oneself, and as to when it now is. We need these possibilities de se et nunc because the

[5]I follow Peter Unger, Ignorance: A Case for Skepticism (New York: Oxford University Press, 1975). But I shall not let him lead me into scepticism.

[6]See Robert Stalnaker. Inquiry (Cambridge, MA: MIT Press, 1984) pp. 59–99.

propositions that may be known include propositions de se et nunc.[7] Not only do I know that there are hands in this world somewhere and somewhen. I know that I have hands. or anyway I have them now. Such propositions aren't just made true or made false by the whole world once and for all. They are true for some of us and not for others, or true at some times and not others, or both.

Further, we cannot limit ourselves to "real" possibilities that conform to the actual laws of nature, and maybe also to actual past history. For propositions about laws and history are contingent, and mayor may not be known.

Neither can we limit ourselves to "epistemic" possibilities for S – possibilities that S does not know not to obtain. That would drain our definition of content. Assume only that knowledge is closed under strict implication. . (We shall consider the merits of this assumption later.) Remember that we are not distinguishing between equivalent propositions. Then knowledge of a conjunction is equivalent to knowledge of every conjunct. P is the conjunction of all propositions not- W, where W is a possibility in which not-P. That suffices to yield an equivalence: S knows that P iff, for every possibility W in which not-P, S knows that not-W. Contraposing and cancelling a double negation: iff every possibility which S does not know not to obtain is one in which P. For short: iff P holds throughout S's epistemic possibilities. Yet to get this far, we need no substantive definition of knowledge at all! To turn this into a substantive definition, in fact the very definition we gave before, we need to say one more thing: S's epistemic possibilities are just those possibilities that are uneliminated by S's evidence.

So, next, we need to say what it means for a possibility to be eliminated or not. Here I say that the uneliminated possibilities are those in which the subject's entire perceptual experience and memory are just as they actually are. There is one possibility that actually obtains (for the subject and at the time in question); call it actuality. Then a possibility W is uneliminated iff the subject's perceptual experience and memory in W exactly match his perceptual experience and memory in actuality. (If you want to include other alleged forms of basic evidence, such as the evidence of our extrasensory faculties, or an innate disposition to believe in God, be my guest. If they exist, they should be included. If not, no harm done if we have included them conditionally.)

Note well that we do not need the "pure sense-datum language" and the "incorrigible protocol statements" that for so long bedevilled foundationalist epistemology. It matters not at all whether there are words to capture the subject's perceptual and memory evidence, nothing more and nothing less. If there are such words, it matters not at all whether the subject can hit upon them. The given does not consist of basic axioms to serve as premises in subsequent arguments. Rather, it consists of a match between possibilities.

[7]See my 'Attitudes De Dicta and De Se', The Philosophical Review 88 (1979) pp. 513–543; and R. M. Chisholm, "The Indirect Reflexive" in C. Diamond and J. Teichman (eds.), Intention and Intentionality: Essays in Honour of G. E. M. Anscomhe (Brighton: Harvester, 1979).

When perceptual experience E (or memory) eliminates a possibility W. that is not because the propositional content of the experience conflicts with W. (Not even if it is the narrow content.) The propositional content of our experience could, after all, be false. Rather, it is the existence of the experience that conflicts with W: W is a possibility in which the subject is not having experience E. Else we would need to tell some fishy story of how the experience has some sort of infallible, ineffable, purely phenomenal propositional content. . . . Who needs that? Let E have propositional content P. Suppose even – something I take to be an open question – that E is, in some sense, fully characterized by P. Then I say that E eliminates W iff W is a possibility in which the subject's experience or memory has content different from P. I do not say that E eliminates W iff W is a possibility in which P is false.

Maybe not every kind of sense perception yields experience; maybe, for instance, the kinaesthetic sense yields not its own distinctive sort of sense experience but only spontaneous judgements about the position of one's limbs. If this is true, then the thing to say is that kinaesthetic evidence eliminates all possibilities except those that exactly resemble actuality with respect to the subject's spontaneous kinaesthetic judgements. In saying this, we would treat kinaesthetic evidence more on the model of memory than on the model of more typical senses.

Finally, we must attend to the word "every." What does it mean to say that every possibility in which not-P is eliminated? An idiom of quantification, like "every," is normally restricted to some limited domain. If I say that every glass is empty, so it's time for another round, doubtless I and my audience are ignoring most of all the glasses there are in the whole wide world throughout all of time. They are outside the domain. They are irrelevant to the truth of what was said.

Likewise, if I say that every uneliminated possibility is one in which P, or words to that effect, I am doubtless ignoring some of all the uneliminated alternative possibilities that there are. They are outside the domain, they are irrelevant to the truth of what was said.

But, of course, I am not entitled to ignore just any possibility I please. Else true ascriptions of knowledge, whether to myself or to others, would be cheap indeed. I may properly ignore some uneliminated possibilities; I may not properly ignore others. Our definition of knowledge requires a sotto voce proviso. S knows that P iff S's evidence eliminates every possibility in which not-P – Psst! – except for those possibilities that we are properly ignoring.

Unger suggests an instructive parallel.[8] Just as P is known iff there are no uneliminated possibilities of error, so likewise a surface is flat iff there are no bumps on it. We must add the proviso: Psst! – except for those bumps that we are properly ignoring. Else we will conclude, absurdly, that nothing is flat. (Simplify by ignoring departures from flatness that consist of gentle curvature.) We can restate the definition. Say that we presuppose proposition Q iff we ignore all possibilities

[8]Peter Unger, Ignorance, chapter II. I discuss the case, and briefly foreshadow the present paper, in my "Scorekeeping in a Language Game," Journal of Philosophical Logic 8 (1979) pp. 339–359, esp. pp. 353–355.

in which not-Q. To close the circle: we ignore just those possibilities that falsify our presuppositions. Proper presupposition corresponds, of course, to proper ignoring. Then S knows that P iff S's evidence eliminates every possibility in which not-P – Psst! – except for those possibilities that conflict with our proper presuppositions.[9]

The rest of (modal) epistemology examines the sotto voce proviso. It asks: what may we properly presuppose in our ascriptions of knowledge? Which of all the uneliminated alternative possibilities may not properly be ignored? Which ones are the ""relevant alternatives"? – relevant, that is, to what the subject does and doesn't know?[10] In reply, we can list several rules.[11] We begin with three prohibitions: rules to tell us what possibilities we may not properly ignore.

First, there is the Rule of Actuality. The possibility that actually obtains is never properly ignored; actuality is always a relevant alternative; nothing false may properly be presupposed. It follows that only what is true is known, wherefore we did not have to include truth in our definition of knowledge. The rule is "externalist" – the subject himself may not be able to tell what is properly ignored. In judging which of his ignorings are proper, hence what he knows, we judge his success in knowing-not how well he tried.

When the Rule of Actuality tells us that actuality may never be properly ignored, we can ask: whose actuality? Ours, when we ascribe knowledge or ignorance to others? Or the subject's? In simple cases, the question is silly. (In fact, it sounds like the sort of pernicious nonsense we would expect from someone who mixes up what is true with what is believed.) There is just one actual world, we the ascribers live in that world, the subject lives there too, so the subject's actuality is the same as ours.

But there are other cases, less simple, in which the question makes perfect sense and needs an answer. Someone mayor may not know who he is; someone may or may not know what time it is. Therefore I insisted that the propositions that may be known must include propositions de se et nunc; and likewise that the possibilities that may be eliminated or ignored must include possibilities de se et nunc. Now we have a good sense in which the subject's actuality may be different from ours. I ask today what Fred knew yesterday. In particular, did he then know who he was? Did he know what day it was? Fred's actuality is the possibility de se et nunc of being

[9]See Robert Stalnaker, "Presuppositions," Journal of Philosophical Logic 2 (1973) pp. 447–457; and "Pragmatic Presuppositions" in Milton Munitz and Peter Unger (eds.), Semantics and Philosophy (New York: New York University Press. 1974). See also my "Score keeping in a Language Game." The definition restated in terms of presupposition resembles the treatment of knowledge in Kenneth S. Ferguson, Philosophical Scepticism (Cornell University doctoral dissertation, 1980).

[10]See Fred Dretske, "Epistemic Operators," The Journal of Philosophy 67 (1970) pp. 1007–1022, and "The Pragmatic Dimension of Knowledge," Philosophical Studies 40 (1981) pp. 363–378; Alvin Goldman, "Discrimination and Perceptual Knowledge:' The Journal of Philosophy 73 (1976) pp. 771–791; G. C. Stine, "Skepticism. Relevant Alternatives, and Deductive Closure," Philosophical Studies 29 (1976) pp. 249–261: and Stewart Cohen. "How to be A Fallibilist," Philosophical Perspectives 2 (1988) pp. 91–123.

[11]Some of them, but only some, taken from the authors just cited.

Fred on September 19th at such-and-such possible world; whereas my actuality is the possibility de se et nunc of being David on September 20th at such-and-such world. So far as the world goes, there is no difference: Fred and I are worldmates, his actual world is the same as mine. But when we build subject and time into the possibilities de se et nunc, then his actuality yesterday does indeed differ from mine today.

What is more, we sometimes have occasion to ascribe knowledge to those who are off at other possible worlds. I didn't read the newspaper yesterday. What would I have known if I had read it? More than I do in fact know. (More and less: I do in fact know that I left the newspaper unread, but if I had read it, I would not have known that I had left it unread.) I-who-did-not-read-the-newspaper am here at this world, ascribing knowledge and ignorance. The subject to whom I am ascribing that knowledge and ignorance, namely I-as-I-would-have-been-if-I-had-read-the-newspaper, is at a different world. The worlds differ in respect at least of a reading of the newspaper. Thus the ascriber's actual world is not the same as the subject's. (I myself think that the ascriber and the subject are two different people: the subject is the ascriber's otherworldly counterpart. But even if you think the subject and the ascriber are the same identical person, you must still grant that this person's actuality qua subject differs from his actuality qua ascriber.)

Or suppose we ask modal questions about the subject: what must he have known, what might he have known? Again we are considering the subject as he is not here, but off at other possible worlds. Likewise if we ask questions about knowledge of knowledge: what does he (or what do we) know that he knows? So the question "whose actuality?" is not a silly question after all. And when the question matters, as it does in the cases just considered, the right answer is that it is the subject's actuality, not the ascriber's, that never can be properly ignored.

Next, there is the Rule of Belief. A possibility that the subject believes to obtain is not properly ignored, whether or not he is right to so believe. Neither is one that he ought to believe to obtain – one that evidence and arguments justify him in believing – whether or not he does so believe.

That is rough. Since belief admits of degree, and since some possibilities are more specific than others, we ought to reformulate the rule in terms of degree of belief, compared to a standard set by the unspecificity of the possibility in question. A possibility may not be properly ignored if the subject gives it, or ought to give it, a degree of belief that is sufficiently high, and high not just because the possibility in question is unspecific.

How high is "sufficiently high"? That may depend on how much is at stake. When error would be especially disastrous, few possibilities may be properly ignored. Then even quite a low degree of belief may be "sufficiently high" to bring the Rule of Belief into play. The jurors know that the accused is guilty only if his guilt has been proved beyond reasonable doubt.[12]

[12]Instead of complicating the Rules of Belief as I have just done. 1 might equivalently have introduced a separate Rule of High Stakes saying that when error would be especially disastrous. few possibilities are properly ignored.

Yet even when the stakes are high, some possibilities still may be properly ignored. Disastrous though it would be to convict an innocent man, still the jurors may properly ignore the possibility that it was the dog, marvellously well-trained, that fired the fatal shot. And, unless they are ignoring other alternatives more relevant than that, they may rightly be said to know that the accused is guilty as charged. Yet if there had been reason to give the dog hypothesis a slightly less negligible degree of belief – if the world's greatest dog-trainer had been the victim's mortal enemy – then the alternative would be relevant after all.

This is the only place where belief and justification enter my story. As already noted, I allow justified true belief without knowledge, as in the case of your belief that you will lose the lottery. I allow knowledge without justification, in the cases of face recognition and chicken sexing. I even allow knowledge without belief, as in the case of the timid student who knows the answer but has no confidence that he has it right, and so does not believe what he knows.[13] Therefore any proposed converse to the Rule of Belief should be rejected. A possibility that the subject does not believe to a sufficient degree, and ought not to believe to a sufficient degree, may nevertheless be a relevant alternative and not properly ignored.

Next, there is the Rule of Resemblance. Suppose one possibility saliently resembles another. Then if one of them may not be properly ignored, neither may the other. (Or rather, we should say that if one of them may not properly be ignored in virtue of rules other than this rule, then neither may the other. Else nothing could be properly ignored; because enough little steps of resemblance can take us from anywhere to anywhere.) Or suppose one possibility saliently resembles two or more others, one in one respect and another in another, and suppose that each of these may not properly be ignored (in virtue of rules other than this rule). Then these resemblances may have an additive effect, doing more together than any one of them would separately.

We must apply the Rule of Resemblance with care. Actuality is a possibility uneliminated by the subject's evidence. Any other possibility W that is likewise uneliminated by the subject's evidence thereby resembles actuality in one salient respect: namely, in respect of the subject's evidence. That will be so even if W is in other respects very dissimilar to actuality – even if, for instance, it is a possibility in which the subject is radically deceived by a demon. Plainly, we dare not apply the Rules of Actuality and Resemblance to conclude that any such W is a relevant alternative – that would be capitulation to scepticism. The Rule of Resemblance was never meant to apply to this resemblance! We seem to have an ad hoc exception to the Rule, though one that makes good sense in view of the function of attributions of knowledge. What would be better, though, would be to find a way to reformulate the Rule so as to get the needed exception without ad hocery. I do not know how to do this.

[13] A. D. Woozley, "Knowing and Not Knowing:' Proceedings of the Aristotelian Society 53 (1953) pp. 151–172: Colin Radford. "Knowledge – by Examples," Analysis 27 (1966) pp. 1–11.

It is the Rule of Resemblance that explains why you do not know that you will lose the lottery, no matter what the odds are against you and no matter how sure you should therefore be that you will lose. For every ticket, there is the possibility that it will win. These possibilities are saliently similar to one another: so either everyone of them may be properly ignored, or else none may. But one of them may not properly be ignored: the one that actually obtains.

The Rule of Resemblance also is the rule that solves the Gettier problems: other cases of justified true belief that are not knowledge.[14]

(1) I think that Nogot owns a Ford, because I have seen him driving one; but unbeknownst to me he does not own the Ford he drives, or any other Ford. Unbeknownst to me, Havit does own a Ford, though I have no reason to think so because he never drives it, and in fact I have often seen him taking the tram. My justified true belief is that one of the two owns a Ford. But I do not know it; I am right by accident. Diagnosis: I do not know, because I have not eliminated the possibility that Nogot drives a Ford he does not own whereas Havit neither drives nor owns a car. This possibility may not properly be ignored. Because, first. actuality may not properly be ignored; and, second. this possibility saliently resembles actuality. It resembles actuality perfectly so far as Nogot is concerned; and it resembles actuality well so far as Havit is concerned, since it matches actuality both with respect to Havit's carless habits and with respect to the general correlation between carless habits and carlessness. In addition, this possibility saliently resembles a third possibility: one in which Nogot drives a Ford he owns while Havit neither drives nor owns a car. This third possibility may not properly be ignored, because of the degree to which it is believed. This time, the resemblance is perfect so far as Havit is concerned, rather good so far as Nogot is concerned.

(2) The stopped clock is right twice a day. It says 4:39. as it has done for weeks. I look at it at 4:39; by luck I pick up a true belief. I have ignored the uneliminated possibility that I looked at it at 4:22 while it was stopped saying 4:39. That possibility was not properly ignored. It resembles actuality perfectly so far as the stopped clock goes.

(3) Unbeknownst to me, I am travelling in the land of the bogus barns; but my eye falls on one of the few real ones. I don't know that I am seeing a barn, because I may not properly ignore the possibility that I am seeing yet another of the

[14]See Edmund Gettier, "Is lustified True Belief Knowledge?," Analysis 23 (1963) pp. 121–123. Diagnoses have varied widely. The four examples below come from: (1) Keith Lehrer and Thomas Paxson Jr., "Knowledge: Undefeated True Belief," The Journal of Philosophy 66 (1969) pp. 225–237; (2) Bertrand Russell, Human Knowledge: Its Scope and Limits (London: Allen and Unwin, 1948) p. 154; (3) Alvin Goldman, "Discrimination and Perceptual Knowledge," op. cit.; (4) Gilbert Harman, Thought (Princeton. N1: Princeton University Press, 1973) p. 143.

Though the lottery problem is another case of justified true belief without knowledge, it is not normally counted among the Gettier problems. It is interesting to find that it yields to the same remedy.

abundant bogus barns. This possibility saliently resembles actuality in respect of the abundance of bogus barns, and the scarcity of real ones, hereabouts.

(4) Donald is in San Francisco, just as I have every reason to think he is. But, bent on deception, he is writing me letters and having them posted to me by his accomplice in Italy. If I had seen the phoney letters, with their Italian stamps and postmarks, I would have concluded that Donald was in Italy. Luckily, I have not yet seen any of them. I ignore the uneliminated possibility that Donald has gone to Italy and is sending me letters from there. But this possibility is not properly ignored, because it resembles actuality both with respect to the fact that the letters are coming to me from Italy and with respect to the fact that those letters come, ultimately, from Donald. So I don't know that Donald is in San Francisco.

Next, there is the Rule of Reliability. This time, we have a presumptive rule about what may be properly ignored; and it is by means of this rule that we capture what is right about causal or reliabilist theories of knowing. Consider processes whereby information is transmitted to us: perception, memory, and testimony. These processes are fairly reliable.[15] Within limits, we are entitled to take them for granted. We may properly presuppose that they work without a glitch in the case under consideration. Defeasibly – very defeasibly! – a possibility in which they fail may properly be ignored.

My visual experience, for instance, depends causally on the scene before my eyes, and what I believe about the scene before my eyes depends in turn on my visual experience. Each dependence covers a wide and varied range of alternatives.[16] Of course, it is possible to hallucinate – even to hallucinate in such a way that all my perceptual experience and memory would be just as they actually are. That possibility never can be eliminated. But it can be ignored. And if it is properly ignored – as it mostly is – then vision gives me knowledge. Sometimes, though, the possibility of hallucination is n properly ignored; for sometimes we really do hallucinate. The Rule of Reliability may be defeated by the Rule of Actuality. Or it may be defeated by the Rules of Actuality and of Resemblance working together, in a Gettier problem: if I am not hallucinating, but unbeknownst to me I live in a world where people mostly do hallucinate and I myself have only narrowly escaped, then the uneliminated possibility of hallucination is too close to actuality to be properly ignored.

We do not, of course, presuppose that nowhere ever is there a failure of, say, vision. The general presupposition that vision is reliable consists, rather, of a

[15]See Alvin Goldman, "A Causal Theory of Knowing," The Journal of Philosophy 64 (1967) pp. 357–372; D. M. Armstrong, Belief, Truth and Knowledge (Cambridge: Cambridge University Press, 1973).

[16]See my "Veridical Hallucination and Prosthetic Vision," Australasian Journal of Philosophy 58 (1980) pp. 239–249. John Bigelow has proposed to model knowledge-delivering processes generally on those found in vision.

standing disposition to presuppose, concerning whatever particular case may be under consideration, that we have no failure in that case.

In similar fashion, we have two permissive Rules of Method. We are entitled to presuppose – again, very defeasibly – that a sample is representative; and that the best explanation of our evidence is the true explanation. That is, we are entitled properly to ignore possible failures in these two standard methods of nondeductive inference. Again, the general rule consists of a standing disposition to presuppose reliability in whatever particular case may come before us.

Yet another permissive rule is the Rule of Conservatism. Suppose that those around us normally do ignore certain possibilities, and it is common knowledge that they do. (They do, they expect each other to, they expect each other to expect each other to, ...) Then – again, very defeasibly! – these generally ignored possibilities may properly be ignored. We are permitted, defeasibly, to adopt the usual and mutually expected presuppositions of those around us.

(It is unclear whether we need all four of these permissive rules. Some might be subsumed under others. Perhaps our habits of treating samples as representative, and of inferring to the best explanation, might count as normally reliable processes of transmission of information. Or perhaps we might subsume the Rule of Reliability under the Rule of Conservatism, on the ground that the reliable processes whereby we gain knowledge are familiar, are generally relied upon, and so are generally presupposed to be normally reliable. Then the only extra work done by the Rule of Reliability would be to cover less familiar – and merely hypothetical? – reliable processes, such as processes that relied on extrasensory faculties. Likewise, mutatis mutandis. we might subsume the Rules of Method under the Rule of Conservatism. Or we might instead think to subsume the Rule of Conservatism under the Rule of Reliability, on the ground that what is generally presupposed tends for the most part to be true, and the reliable processes whereby this is so are covered already by the Rule of Reliability. Better redundancy than incompleteness, though. So. leaving the question of redundancy open, I list all four rules.)

Our final rule is the Rule of Attention. But it is more a triviality than a rule. When we say that a possibility is properly ignored, we mean exactly that; we do not mean that it could have been properly ignored. Accordingly, a possibility not ignored at all is ipso facto not properly ignored. What is and what is not being ignored is a feature of the particular conversational context. No matter how far-fetched a certain possibility may be, no matter how properly we might have ignored it in some other context, if in this context we are not in fact ignoring it but attending to it, then for us now it is a relevant alternative. It is in the contextually determined domain. If it is an uneliminated possibility in which not-P, then it will do as a counter-example to the claim that P holds in every possibility left uneliminated by S's evidence. That is, it will do as a counter-example to the claim that S knows that P.

Do some epistemology. Let your fantasies rip. Find uneliminated possibilities of error everywhere. Now that you are attending to them. just as I told you to, you are no longer ignoring them, properly or otherwise. So you have landed in a context with an enormously rich domain of potential counter-examples to ascriptions of knowledge. In such an extraordinary context, with such a rich domain, it never

can happen (well, hardly ever) that an ascription of knowledge is true. Not an ascription of knowledge to yourself (either to your present self or to your earlier self, untainted by epistemology); and not an ascription of knowledge to others. That is how epistemology destroys knowledge. But it does so only temporarily.

The pastime of epistemology does not plunge us forevermore into its special context. We can still do a lot of proper ignoring, a lot of knowing, and a lot of true ascribing of knowledge to ourselves and others, the rest of the time.

What is epistemology all about? The epistemology we've just been doing, at any rate, soon became an investigation of the ignoring of possibilities. But to investigate the ignoring of them was ipso facto not to ignore them. Unless this investigation of ours was an altogether atypical sample of epistemology, it will be inevitable that epistemology must destroy knowledge. That is how knowledge is elusive. Examine it, and straightway it vanishes.

Is resistance useless? If you bring some hitherto ignored possibility to our attention, then straightway we are not ignoring it at all, so a fortiori we are not properly ignoring it. How can this alteration of our conversational state be undone? If you are persistent, perhaps it cannot be undone – at least not so long as you are around. Even if we go off and play backgammon, and afterward start our conversation afresh, you might turn up and call our attention to it all over again.

But maybe you called attention to the hitherto ignored possibility by mistake. You only suggested that we ought to suspect the butler because you mistakenly thought him to have a criminal record. Now that you know he does not – that was the previous butler – you wish you had not mentioned him at all. You know as well as we do that continued attention to the possibility you brought up impedes our shared conversational purposes. Indeed, it may be common knowledge between you and us that we would all prefer it if this possibility could be dismissed from our attention. In that case we might quickly strike a tacit agreement to speak just as if we were ignoring it; and after just a little of that, doubtless it really would be ignored.

Sometimes our conversational purposes are not altogether shared, and it is a matter of conflict whether attention to some far-fetched possibility would advance them or impede them. What if some farfetched possibility is called to our attention not by a sceptical philosopher, but by counsel for the defence? We of the jury may wish to ignore it, and wish it had not been mentioned. If we ignored it now, we would bend the rules of cooperative conversation; but we may have good reason to do exactly that. (After all, what matters most to us as jurors is not whether we can truly be said to know; what really matters is what we should believe to what degree, and whether or not we should vote to convict.) We would ignore the far-fetched possibility if we could – but can we? Perhaps at first our attempted ignoring would be make-believe ignoring, or self-deceptive ignoring; later, perhaps, it might ripen into genuine ignoring. But in the meantime, do we know? There may be no definite answer. We are bending the rules, and our practices of context-dependent attributions of knowledge were made for contexts with the rules unbent.

If you are still a contented fallibilist, despite my plea to hear the sceptical argument afresh, you will probably be discontented with the Rule of Attention. You will begrudge the sceptic even his very temporary victory. You will claim the right to

resist his argument not only in everyday contexts, but even in those peculiar contexts in which he (or some other epistemologist) busily calls your attention to farfetched possibilities of error. Further, you will claim the right to resist without having to bend any rules of cooperative conversation. I said that the Rule of Attention was a triviality: that which is not ignored at all is not properly ignored. But the Rule was trivial only because of how I had already chosen to state the sotto voce proviso. So you, the contented fallibilist, will think it ought to have been stated differently. Thus, perhaps: "Psst! – except for those possibilities we could properly have ignored". And then you will insist that those far-fetched possibilities of error that we attend to at the behest of the sceptic are nevertheless possibilities we could properly have ignored. You will say that no amount of attention can, by itself, turn them into relevant alternatives.

If you say this, we have reached a standoff. I started with a puzzle: how can it be, when his conclusion is so silly, that the sceptic's argument is so irresistible? My Rule of Attention, and the version of the proviso that made that Rule trivial, were built to explain how the sceptic manages to sway us – why his argument seems irresistible, however temporarily. If you continue to find it eminently resistible in all contexts, you have no need of any such explanation. We just disagree about the explanandum phenomenon.

I say S knows that P iff P holds in every possibility left uneliminated by S's evidence – Psst! – except for those possibilities that we are properly ignoring. "We" means: the speaker and hearers of a given context; that is, those of us who are discussing S's knowledge together. It is our ignorings, not S's own ignorings, that matter to what we can truly say about S's knowledge. When we are talking about our own knowledge or ignorance, as epistemologists so often do, this is a distinction without a difference. But what if we are talking about someone else?

Suppose we are detectives; the crucial question for our solution of the crime is whether S already knew, when he bought the gun, that he was vulnerable to blackmail. We conclude that he did. We ignore various far-fetched possibilities, as hard-headed detectives should. But S does not ignore them. S is by profession a sceptical epistemologist. He never ignores much of anything. If it is our own ignorings that matter to the truth of our conclusion, we may well be right that S already knew. But if it is S's ignorings that matter, then we are wrong, because S never knew much of anything. I say we may well be right; so it is our own ignorings that matter, not S's.

But suppose instead that we are epistemologists considering what S knows. If we are well-informed about S (or if we are considering a well-enough specified hypothetical case), then if S attends to a certain possibility, we attend to S's attending to it. But to attend to 5's attending to it is ipso facto to attend to it ourselves. In that case, unlike the case of the detectives. the possibilities we are properly ignoring must be among the possibilities that S himself ignores. We may ignore fewer possibilities than S does, but not more.

Even if S himself is neither sceptical nor an epistemologist, he may yet be clever at thinking up farfetched possibilities that are uneliminated by his evidence. Then again, we well-informed epistemologists who ask what S knows will have

to attend to the possibilities that S thinks up. Even if S's idle cleverness does not lead S himself to draw sceptical conclusions, it nevertheless limits the knowledge that we can truly ascribe to him when attentive to his state of mind. More simply: his cleverness limits his knowledge. He would have known more, had he been less imaginative.[17]

Do I claim you can know P just by presupposing it?! Do I claim you can know that a possibility W does not obtain just by ignoring it? Is that not what my analysis implies, provided that the presupposing and the ignoring are proper? Well, yes. And yet I do not claim it. Or rather, I do not claim it for any specified P or W. I have to grant, in general, that knowledge just by presupposing and ignoring is knowledge; but it is an especially elusive sort of knowledge, and consequently it is an unclaimable sort of knowledge. You do not even have to practise epistemology to make it vanish. Simply mentioning any particular case of this knowledge, aloud or even in silent thought, is a way to attend to the hitherto ignored possibility. and thereby render it no longer ignored, and thereby create a context in which it is no longer true to ascribe the knowledge in question to yourself or others. So. just as we should think, presuppositions alone are not a basis on which to claim knowledge.

In general, when S knows that P some of the possibilities in which not-P are eliminated by S's evidence and others of them are properly ignored. There are some that can be eliminated, but cannot properly be ignored. For instance, when I look around the study without seeing Possum the cat, I thereby eliminate various possibilities in which Possum is in the study; but had those possibilities not been eliminated, they could not properly have been ignored. And there are other possibilities that never can be eliminated, but can properly be ignored. For instance, the possibility that Possum is on the desk but has been made invisible by a deceiving demon falls normally into this class (though not when I attend to it in the special context of epistemology).

There is a third class: not-P possibilities that might either be eliminated or ignored. Take the farfetched possibility that Possum has somehow managed to get into a closed drawer of the desk-maybe he jumped in when it was open, then I closed it without noticing him. That possibility could be eliminated by opening the drawer and making a thorough examination. But if uneliminated, it may nevertheless be ignored, and in many contexts that ignoring would be proper. If I look all around the study, but without checking the closed drawers of the desk, I may truly be said to know that Possum is not in the study – Dr at any rate, there are many contexts in which that may truly be said. But if I did check all the closed drawers, then I would know better that Possum is not in the study. My knowledge would be better in the

[17]See Catherine Elgin, "The Epistemic Efficacy of Stupidity," Synthese 74 (1988) pp. 297–311. The "efficacy" takes many forms; some to do with knowledge (under various rival analyses), some to do with justified belief. See also Michael Williams, Unnatural Doubts: Epistemological Realism and the Basis of Scepticism (Oxford: Blackwell, 1991) pp. 352–355, on the instability of knowledge under reflection.

second case because it would rest more on the elimination of not-P possibilities, less on the ignoring of them.[18,19]

Better knowledge is more stable knowledge: it stands more chance of surviving a shift of attention in which we begin to attend to some of the possibilities formerly ignored. If, in our new shifted context, we ask what knowledge we may truly ascribe to our earlier selves, we may find that only the better knowledge of our earlier selves still deserves the name. And yet, if our former ignorings were proper at the time, even the worse knowledge of our earlier selves could truly have been called knowledge in the former context.

Never – well, hardly ever – does our knowledge rest entirely on elimination and not at all on ignoring. So hardly ever is it quite as good as we might wish. To that extent. the lesson of scepticism is right – and right permanently, not just in the temporary and special context of epistemology.[20]

What is it all for? Why have a notion of knowledge that works in the way I described? (Not a compulsory question. Enough to observe that short-cuts – like satisficing, like having indeterminate degrees of belief – that we resort to because we are not smart enough to live up to really high, perfectly Bayesian, standards of rationality. You cannot maintain a record of exactly which possibilities you have eliminated so far, much as you might like to. It is easier to keep track of which possibilities you have eliminated if you – Psst! – ignore many of all the possibilities there are. And besides, it is easier to list some of the propositions that are true in all the uneliminated, unignored possibilities than it is to find propositions that are true in all and only the uneliminated, unignored possibilities.

If you doubt that the word "know" bears any real load in science or in metaphysics, I partly agree. The serious business of science has to do not with knowledge per se; but rather, with the elimination of possibilities through the evidence of perception, memory, etc., and with the changes that one's belief system would (or might or should) undergo under the impact of such eliminations. Ascriptions of

[18]Mixed cases are possible: Fred properly ignores the possibility W1 which Ted eliminates; however, Ted properly ignores the possibility W2 which Fred eliminates. Ted has looked in all the desk drawers but not the file drawers, whereas Fred has checked the file drawers but not the desk. Fred's knowledge that Possum is not in the study is better in one way, Ted's is better in another.

[19]To say truly that X is known, I must be properly ignoring any uneliminated possibilities in which not-X; whereas to say truly that Y is better known than X, I must be attending to some such possibilities. So I cannot say both in a single context. If I say "X is known, but Y is better known," the context changes in mid-sentence: some previously ignored possibilities must stop being ignored. That can happen easily. Saying it the other way around – "Y is better known than X, but even X is known" – is harder, because we must suddenly start to ignore previously unignored possibilities. That cannot be done, really; but we could bend the rules and make believe we had done it, and no doubt we would be understood well enough. Saying "X is flat, but Y is flatter" (that is, "X has no bumps at all, but Y has even fewer or smaller bumps") is a parallel case. And again, "Y is flatter, but even X is flat" sounds clearly worse – but not altogether hopeless.

[20]Thanks here to Stephen Hetherington. While his own views about better and worse knowledge are situated within an analysis of knowledge quite unlike mine, they withstand transplantation.

knowledge to yourself or others are a very sloppy way of conveying very incomplete information about the elimination of possibilities. It is as if you had said:

> The possibilities eliminated, whatever else they may also include, at least include all the not-P possibilities; or anyway, all of those except for some we are presumably prepared to ignore just at the moment.

The only excuse for giving information about what really matters in such a sloppy way is that at least it is easy and quick! But it is easy and quick; whereas giving full and precise information about which possibilities have been eliminated seems to be extremely difficult, as witness the futile search for a "pure observation language." If I am right about how ascriptions of knowledge work, they are a handy but humble approximation. They may yet be indispensable in practice, in the same way that other handy and humble approximations are.

If we analyse knowledge as a modality, as we have done, we cannot escape the conclusion that knowledge is closed under (strict) implication.[21] Dretske has denied that knowledge is closed under implication; further, he has diagnosed closure as the fallacy that drives arguments for scepticism. As follows: the proposition that I have hands implies that I am not a handless being, and a fortiori that I am not a handless being deceived by a demon into thinking that I have hands. So, by the closure principle, the proposition that I know I have hands implies that I know that I am not handless and deceived. But I don't know that I am not handless and deceived – for how can I eliminate that possibility? So, by modus tollens, I don't know that I have hands. Dretske's advice is to resist scepticism by denying closure. He says that although having hands does imply not being handless and deceived, yet knowing that I have hands does not imply knowing that I am not handless and deceived. I do know the former, I do not know the latter.[22]

What Dretske says is close to right, but not quite. Knowledge is closed under implication. Knowing that I have hands does imply knowing that I am not handless and deceived. Implication preserves truth – that is, it preserves truth in any given, fixed context. But if we switch contexts midway, all bets are off. I say (1) pigs fly; (2) what I just said had fewer than three syllables (true); (3) what I just said had fewer than four syllables (false). So "less than three" does not imply "less than four"? No! The context switched midway, the semantic value of the context-dependent

[21] A proof-theoretic version of this closure principle is common to all "normal" modal logics: if the logic validates an inference from zero or more premises to a conclusion, then also it validates the inference obtained by prefixing the necessity operator to each premise and to the conclusion. Further, this rule is all we need to take us from classical sentential logic to the least normal modal logic. See Brian Chellas, Modal Logic: An Introduction (Cambridge: Cambridge University Press, 1980) p. 114.

[22] Dretske, "Epistemic Operators." My reply follows the lead of Stine, "Skepticism, Relevant Alternatives, and Deductive Closure," op. cit.; and (more closely) Cohen, "How to be a Fallibilist," op. cit.

phrase "what I just said" switched with it. Likewise in the sceptical argument the context switched midway, and the semantic value of the context-dependent word "know" switched with it. The premise "I know that I have hands" was true in its everyday context, where the possibility of deceiving demons was properly ignored. The mention of that very possibility switched the context midway. The conclusion "I know that I am not handless and deceived" was false in its context, because that was a context in which the possibility of deceiving demons was being mentioned, hence was not being ignored, hence was not being properly ignored. Dretske gets the phenomenon right, and I think he gets the diagnosis of scepticism right; it is just that he misclassifies what he sees. He thinks it is a phenomenon of logic, when really it is a phenomenon of pragmatics. Closure, rightly understood, survives the test. If we evaluate the conclusion for truth not with respect to the context in which it was uttered, but instead with respect to the different context in which the premise was uttered, then truth is preserved. And if, per impossible, the conclusion could have been said in the same unchanged context as the premise, truth would have been preserved.

A problem due to Saul Kripke turns upon the closure of knowledge under implication. P implies that any evidence against P is misleading. So, by closure, whenever you know that P. you know that any evidence against P is misleading. And if you know that evidence is misleading, you should pay it no heed. Whenever we know – and we know a lot, remember – we should not heed any evidence tending to suggest that we are wrong, But that is absurd. Shall we dodge the conclusion by denying closure? I think not. Again, I diagnose a change of context. At first, it was stipulated that S knew, whence it followed that S was properly ignoring all possibilities of error. But as the story continues, it turns out that there is evidence on offer that points to some particular possibility of error. Then, by the Rule of Attention, that possibility is no longer properly ignored, either by S himself or by we who are telling the story of S. The advent of that evidence destroys S's knowledge, and thereby destroys S's licence to ignore the evidence lest he be misled.

There is another reason, different from Dretske's. why we might doubt closure. Suppose two or more premises jointly imply a conclusion. Might not someone who is compartmentalized in his thinking – as we all are – know each of the premises but fail to bring them together in a single compartment? Then might he not fail to know the conclusion? Yes; and I would not like to plead idealization-of-rationality as an excuse for ignoring such cases. But I suggest that we might take not the whole compartmentalized thinker, but rather each of his several overlapping compartments, as our "subjects." That would be the obvious remedy if his compartmentalization amounted to a case of multiple personality disorder; but maybe it is right for milder cases as well.[23]

A compartmentalized thinker who indulges in epistemology can destroy his knowledge, yet retain it as well. Imagine two epistemologists on a bushwalk. As they walk, they talk. They mention all manner of far-fetched possibilities of error.

[23] See Stalnaker, Inquiry, pp. 79–99.

By attending to these normally ignored possibilities they destroy the knowledge they normally possess. Yet all the while they know where they are and where they are going! How so? The compartment in charge of philosophical talk attends to far-fetched possibilities of error. The compartment in charge of navigation does not. One compartment loses its knowledge, the other retains its knowledge. And what does the entire compartmentalized thinker know? Not an altogether felicitous question. But if we need an answer, I suppose the best thing to say is that S knows that P iff anyone of S's compartments knows that P. Then we can say what we would offhand want to say: yes, our philosophical bushwalkers still know their whereabouts.

Context-dependence is not limited to the ignoring and non-ignoring of farfetched possibilities. Here is another case. Pity poor Bill! He squanders all his spare cash on the pokies, the races, and the lottery. He will be a wage slave all his days. We know he will never be rich. But if he wins the lottery (if he wins big), then he will be rich. Contrapositively: his never being rich, plus other things we know, imply that he will lose. So, by closure, if we know that he will never be rich, we know that he will lose. But when we discussed the case before, we concluded that we cannot know that he will lose. All the possibilities in which Bill loses and someone else wins saliently resemble the possibility in which Bill wins and the others lose; one of those possibilities is actual; so by the Rules of Actuality and of Resemblance, we may not properly ignore the possibility that Bill wins. But there is a loophole: the resemblance was required to be salient. Salience, as well as ignoring, may vary between contexts. Before, when I was explaining how the Rule of Resemblance applied to lotteries, I saw to it that the resemblance between the many possibilities associated with the many tickets was sufficiently salient. But this time, when we were busy pitying poor Bill for his habits and not for his luck, the resemblance of the many possibilities was not so salient. At that point, the possibility of Bill's winning was properly ignored; so then it was true to say that we knew he would never be rich. Afterward I switched the context. I mentioned the possibility that Bill might win, wherefore that possibility was no longer properly ignored. (Maybe there were two separate reasons why it was no longer properly ignored, because maybe I also made the resemblance between the many possibilities more salient.) It was true at first that we knew that Bill would never be rich. And at that point it was also true that we knew he would lose – but that was only true so long as it remained unsaid! (And maybe unthought as well.) Later, after the change in context, it was no longer true that we knew he would lose. At that point, it was also no longer true that we knew he would never be rich.

But wait. Don't you smell a rat? Haven't I, by my own lights, been saying what cannot be said? (Or whistled either.) If the story I told was true, how have I managed to tell it? In trendyspeak, is there not a problem of reflexivity? Does not my story deconstruct itself?

I said: S knows that P iff S's evidence eliminates every possibility in which not-P – Psst! – except for those possibilities that we are properly ignoring. That "psst" marks an attempt to do the impossible – to mention that which remains unmentioned. I am sure you managed to make believe that I had succeeded. But I could not have done.

And I said that when we do epistemology, and we attend to the proper ignoring of possibilities, we make knowledge vanish. First we do know, then we do not. But I had been doing epistemology when I said that. The uneliminated possibilities were Hot being ignored – not just then. So by what right did I say even that we used to know?[24]

In trying to thread a course between the rock of fallibilism and the whirlpool of scepticism, it may well seem as if I have fallen victim to both at once. For do I not say that there are all those uneliminated possibilities of error? Yet do I not claim that we know a lot? Yet do I not claim that knowledge is, by definition, infallible knowledge?

I did claim all three things. But not all at once! Or if I did claim them all at once, that was an expository shortcut, to be taken with a pinch of salt. To get my message across, I bent the rules. If I tried to whistle what cannot be said, what of it? I relied on the cardinal principle of pragmatics, which overrides every one of the rules I mentioned: interpret the message to make it make sense – to make it consistent, and sensible to say.

When you have context-dependence, ineffability can be trite and unmysterious. Hush! [moment of silence] I might have liked to say, just then, "All of us are silent." It was true. But I could not have said it truly, or whistled it either. For by saying it aloud, or by whistling, I would have rendered it false.

I could have said my say fair and square, bending no rules. It would have been tiresome, but it could have been done. The secret would have been to resort to "semantic ascent." I could have taken great care to distinguish between (1) the language I use when I talk about knowledge, or whatever, and (2) the second language that I use to talk about the semantic and pragmatic workings of the first language. If you want to hear my story told that way, you probably know enough to do the job for yourself. If you can, then my informal presentation has been good enough.

[24]Worse still: by what right can I even say that we used to be in a position to say truly that we knew? Then, we were in a context where we properly ignored certain uneliminated possibilities of error. Now, we are in a context where we no longer ignore them. If now I comment retrospectively upon the truth of what was said then, which context governs: the context now or the context then? I doubt there is any general answer, apart from the usual principle that we should interpret what is said so as to make the message make sense.

Chapter 29
Knowledge and Scepticism

Robert Nozick

You think you are seeing these words, but could you not be hallucinating or dreaming or having your brain stimulated to give you the experience of seeing these marks on paper although no such thing is before you? More extremely, could you not be floating in a tank while super-psychologists stimulate your brain electrochemically to produce exactly the same experiences as you are now having, or even to produce the whole sequence of experiences you have had in your lifetime thus far? If one of these other things was happening, your experience would be exactly the same as it now is. So how can you know none of them is happening? Yet if you do not know these possibilities don't hold, how can you know you are reading this book now? If you do not know you haven't always been floating in the tank at the mercy of the psychologists, how can you know anything-what your name is, who your parents were, where you come from?

The sceptic argues that we do not know what we think we do. Even when he leaves us unconverted, he leaves us confused. Granting that we do know, how *can* we? Given these other possibilities he poses, how is knowledge possible? In answering this question. we do not seek to convince the sceptic, but rather to formulate hypotheses about knowledge and our connection to facts that show how knowledge can exist even given the sceptic's possibilities. These hypotheses must reconcile our belief that we know things with our belief that the sceptical possibilities are logical possibilities.

The sceptical possibilities, and the threats they pose to our knowledge, depend upon our knowing things (if we do) mediately, through or by way of something else. Our thinking or believing that some fact p holds is connected somehow to the

Robert Nozick was deceased at the time of publication.

R. Nozick (deceased)
Harvard University, Boston, MA, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_29

fact that *p*, but is not itself identical with that fact. Intermediate links establish the connection. This leaves room for the possibility of these intermediate stages holding and producing our belief that *p*, without the fact that *p* being at the other end. The intermediate stages arise in a completely different manner, one not involving the fact that *p* although giving rise to the appearance that *p* holds true.

Are the sceptic's possibilities indeed logically possible? Imagine reading a science fiction story in which someone is raised from birth floating in a tank with psychologists stimulating his brain. The story could go on to tell of the person's reactions when he is brought out of the tank, of how the psychologists convince him of what had been happening to him, or how they fail to do so. This story is coherent, there is nothing self-contradictory or otherwise impossible about it. Nor is there anything incoherent in imagining that you are now in this situation, at a time before being taken out of the tank. To ease the transition out, to prepare the way, perhaps the psychologists will give the person in the tank thoughts of whether floating in the tank is possible, or the experience of reading a book that discusses this possibility, even one that discusses their easing his transition. (Free will presents no insuperable problem for this possibility. Perhaps the psychologists caused all your experiences of choice, including the feeling of freely choosing; or perhaps you do freely choose to act while they, cutting the effector circuit, continue the scenario from there.)

Some philosophers have attempted to demonstrate there is no such coherent possibility of this sort. However, for any reasoning that purports to show this sceptical possibility cannot occur, we can imagine the psychologists of our science fiction story feeding it to their tank-subject, along with the (inaccurate) feeling that the reasoning is cogent. So how much trust can be placed in the apparent cogency of an argument to show the sceptical possibility isn't coherent? The sceptic's possibility is a logically coherent one, in tension with the existence of (almost all) knowledge; so we seek a hypothesis to explain how, even given the sceptic's possibilities, knowledge is possible. We may worry that such explanatory hypotheses are ad hoc, but this worry will lessen if they yield other facts as well, fit in with other things we believe, and so forth. Indeed, the theory of knowledge that follows was not developed in order to explain how knowledge is possible. Rather, the motivation was external to epistemology; only after the account of knowledge was developed for another purpose did I notice its consequences for scepticism, for understanding how knowledge is possible. So whatever other defects the explanation might have, it can hardly be called ad hoc.

Knowledge

Our task is to formulate further conditions to go alongside

(1) *p* is true
(2) S believes that *p*.

We would like each condition to be necessary for knowledge, so any case that fails to satisfy it will not be an instance of knowledge. Furthermore, we would like the conditions to be jointly sufficient for knowledge, so any case that satisfies all of them will be an instance of knowledge. We first shall formulate conditions that seem to handle ordinary cases correctly, classifying as knowledge cases which are knowledge, and as non-knowledge cases which are not; then we shall check to see how these conditions handle some difficult cases discussed in the literature.

One plausible suggestion is causal, something like: the fact that p (partially) causes S to believe that p, that is, (2) because (1). But this provides an inhospitable environment for mathematical and ethical knowledge; also there are well-known difficulties in specifying the type of causal connection. If someone floating in a tank oblivious to everything around him is given (by direct electrical and chemical stimulation of the brain) the belief that he is floating in a tank with his brain being stimulated, then even though that fact is part of the cause of his belief, still he does not know that it is true. Let us consider a different third condition:

(3) If p were not true, S would not believe that p.

Throughout this work, let us write the subjunctive "if-then" by an arrow, and the negation of a sentence by prefacing "not-" to it. The above condition thus is rewritten as:

(3) not-$p \rightarrow$ not-(S believes that p).

This subjunctive condition is not unrelated to the causal condition. Often when the fact that p (partially) causes someone to believe that p, the fact also will be causally necessary for his having the belief without the cause, the effect would not occur. In that case, the subjunctive condition (3) also will be satisfied. Yet this condition is not equivalent to the causal condition. For the causal condition will be satisfied in cases of causal overdetermination, where either two sufficient causes of the effect actually operate, or a back-up cause (of the same effect) would operate if the first one didn't; whereas the subjunctive condition need not hold for these cases.[1,2] When the two conditions do agree, causality indicates knowledge because it acts in a manner that makes the subjunctive (3) true.

The subjunctive condition (3) serves to exclude cases of the sort first described by Edward Gettier, such as the following. Two other people are in my office and I am justified on the basis of much evidence in believing the first owns a Ford car; though he (now) does not, the second person (a stranger to me) owns one. I believe truly and justifiably that someone (or other) in my office owns a Ford car, but I do not know someone does. Concluded Gettier, knowledge is not simply justified true belief.

[1] See Hilary Putnam, Reason, Truth and History (Cambridge, 1981), ch. I.

[2] I should note here that I assume bivalence throughout this chapter, and consider only statements that are true if and only if their negations are false.

The following subjunctive, which specifies condition (3) for this Gettier case, is not satisfied: if no one in my office owned a Ford car, I wouldn't believe that someone did. The situation that would obtain if no one in my office owned a Ford is one where the stranger does not (or where he is not in the office); and in that situation I still would believe, as before, that someone in my office does own a Ford, namely, the first person. So the subjunctive condition (3) excludes this Gettier case as a case of knowledge.

The subjunctive condition is powerful and intuitive, not so easy to satisfy, yet not so powerful as to rule out everything as an instance of knowledge. A subjunctive conditional "if p were true, q would be true," $p \rightarrow q$, does not say that p entails q or that it is logically impossible that p yet not-q. It says that in the situation that would obtain if p were true, q also would be true. This point is brought out especially clearly in recent "possible-worlds" accounts of subjunctives: the subjunctive is true when (roughly) in all those worlds in which p holds true that are closest to the actual world, q also is true. (Examine those worlds in which p holds true closest to the actual world, and see if q holds true in all these.) Whether or not q is true in p worlds that are still farther away from the actual world is irrelevant to the truth of the subjunctive. I do not mean to endorse any particular possible-worlds account of subjunctives, nor am I committed to this type of account.[3] I sometimes shall use it, though, when it illustrates points in an especially clear way.

The subjunctive condition (3) also handles nicely cases that cause difficulties for the view that you know that p when you can rule out the relevant alternatives to p in the context. For, as Gail Stine writes, "what makes an alternative relevant in one context and not another? ... if on the basis of visual appearances obtained under optimum conditions while driving through the countryside Henry identifies an object as a barn, normally we say that Henry knows that it is a barn. Let us suppose, however, that unknown to Henry, the region is full of expertly made papier-mache facsimiles of barns. In that case, we would not say that Henry knows that the object is a barn, unless he has evidence against it being a papier-mache facsimile, which is now a relevant alternative. So much is clear, but what if no such facsimiles exist in Henry's surroundings, although they once did? Are either of these circumstances sufficient to make the hypothesis (that it's a papier-mache object) relevant? Probably not, but the situation is not so clear:"[4] Let p be the statement that the object in the field is a (real) barn, and q the one that the object in the field is a papier-mache barn. When papier-mache barns are scattered through the area, if p were false, q would be true or might be. Since in this case (we are supposing) the person still would believe p, the subjunctive

[3] See Robert Stalnaker, "A Theory of Conditionals," in N. Rescher, ed., *Studies in Logical Theory* (Oxford 1968); David Lewis, Counteifactuals (Cambridge 1973); and Jonathan Bennett's critical review of Lewis, "Counterfactuals and Possible Worlds," *Canadian Journal of Philosophy*, 4/2 (Dec. 1974), 381–402. Our purposes require, for the most part, no more than an intuitive understanding of subjunctives.

[4] G. C. Stine, "Skepticism, Relevant Alternatives and Deductive Closure," *Philosophical Studies*, 29 (1976), 252, who attributes the example to Carl Ginet.

(3) not-$p \rightarrow$ not-(S believes that p)

is not satisfied, and so he doesn't know that p. However, when papier-mache barns are or were scattered around another country, even if p were false q wouldn't be true, and so (for all we have been told) the person may well know that p. A hypothesis q contrary to p clearly is relevant when if p weren't true, q would be true; when not-$p \rightarrow q$. It clearly is irrelevant when if p weren't true, q also would not be true; when not-$p \rightarrow$ not-q. The remaining possibility is that neither of these opposed subjunctives holds; q might (or might not) be true if p weren't true. In this case, q also will be relevant, according to an account of knowledge incorporating condition (3) and treating subjunctives along the lines sketched above. Thus, condition (3) handles cases that befuddle the "relevant alternatives" account; though that account can adopt the above subjunctive criterion for when an alternative is relevant, it then becomes merely an alternate and longer way of stating condition (3).

Despite the power and intuitive force of the condition that if p weren't true the person would not believe it, this condition does not (in conjunction with the first two conditions) rule out every problem case. There remains, for example, the case of the person in the tank who is brought to believe, by direct electrical and chemical stimulation of his brain, that he is in the tank and is being brought to believe things in this way; he does not know this is true. However, the subjunctive condition is satisfied: if he weren't floating in the tank, he wouldn't believe he was.

The person in the tank does not know he is there, because his belief is not sensitive to the truth. Although it is caused by the fact that is its content, it is not sensitive to that fact. The operators of the tank could have produced any belief, including the false belief that he wasn't in the tank; if they had, he would have believed that. Perfect sensitivity would involve beliefs and facts varying together. We already have one portion of that variation, subjunctively at least: if p were false he wouldn't believe it. This sensitivity as specified by a subjunctive does not have the belief vary with the truth or falsity of p in all possible situations, merely in the ones that would or might obtain if p were false. The subjunctive condition

(3) not-$p \rightarrow$ not-(S believes that p)

tells us only half the story about how his belief is sensitive to the truth-value of p. It tells us how his belief state is sensitive to p's falsity, but not how it is sensitive to p's truth; it tells us what his belief state would be if p were false, but not what it would be if p were true. To be sure, conditions (1) and (2) tell us that p is true and he does believe it, but it does not follow that his believing p is sensitive to p's being true. This additional sensitivity is given to us by a further subjunctive: if p were true, he would believe it.

(4) $p \rightarrow$ S believes that p.

Not only is p true and S believes it, but if it were true he would believe it. Compare: not only was the photon emitted and did it go to the left, but (it was then true that): if it were emitted it would go to the left. The truth of antecedent and consequent is not alone sufficient for the truth of a subjunctive; (4) says more than

(1) and (2). Thus, we presuppose some (or another) suitable account of subjunctives. According to the suggestion tentatively made above, (4) holds true if not only does he actually truly believe *p*, but in the 'close' worlds where *p* is true, he also believes it. He believes that *p* for some distance out in the *p*. neighbourhood of the actual world; similarly, condition (3) speaks not of the whole not-*p* neighbourhood of the actual world, but only of the first portion of it. (If, as is likely, these explanations do not help, please use your own intuitive understanding of the subjunctives (3) and (4).)

The person in the tank does not satisfy the subjunctive condition (4). Imagine as actual a world in which he is in the tank and is stimulated to believe he is, and consider what subjunctives are true in that world. It is not true of him there that if he were in the tank he would believe it; for in the close world (or situation) to his own where he is in the tank but they don't give him the belief that he is (much less instill the belief that he isn't) he doesn't believe he is in the tank. Of the person actually in the tank and believing it, it is not true to make the further statement that if he were in the tank he would believe it – so he does not know he is in the tank.

The subjunctive condition (4) also handles a case presented by Gilbert Harman.8 The dictator of a country is killed; in their first edition, newspapers print the story, but later all the country's newspapers and other media deny the story, falsely. Everyone who encounters the denial believes it (or does not know what to believe and so suspends judgement). Only one person in the country fails to hear any denial and he continues to believe the truth. He satisfies conditions (1) – (3) (and the causal condition about belief) yet we are reluctant to say he knows the truth. The reason is that if he had heard the denials, he too would have believed them, just like everyone else. His belief is not sensitively tuned to the truth, he doesn't satisfy the condition that if it were true he would believe it. Condition (4) is not satisfied.

There is a pleasing symmetry about how this account of knowledge relates conditions (3) and (4), and connects them to the first two conditions. The account has the following form.

(1)
(2)
(3) not-l $\rightarrow$ not-2
(4) 1 $\rightarrow$ 2

I am not inclined, however, to make too much of this symmetry, for I found also that with other conditions experimented with as a possible fourth condition there was some way to construe the resulting third and fourth conditions as symmetrical answers to some symmetrical looking questions, so that they appeared to arise in parallel fashion from similar questions about the components of true belief.

Symmetry, it seems, is a feature of a mode of presentation, not of the contents presented. A uniform transformation of symmetrical statements can leave the results non-symmetrical. But if symmetry attaches to mode of presentation, how can it possibly be a deep feature of, for instance, laws of nature that they exhibit symmetry? (One of my favourite examples of symmetry is due to Groucho Marx. On his radio programme he spoofed a commercial, and ended. "And if you are not

completely satisfied, return the unused portion of our product and we will return the unused portion of your money.") Still, to present our subject symmetrically makes the connection of knowledge to true belief especially perspicuous. It seems to me that a symmetrical formulation is a sign of our understanding, rather than a mark of truth. If we cannot understand an asymmetry as arising from an underlying symmetry through the operation of a particular factor, we will not understand why that asymmetry exists in that direction. (But do we also need to understand why the underlying asymmetrical factor holds instead of its opposite?)

A person knows that *p* when he not only does truly believe it, but also would truly believe it and wouldn't falsely believe it. He not only actually has a true belief, he subjunctively has one. It is true that *p* and he believes it; if it weren't true he wouldn't believe it, and if it were true he would believe it. To know that *p* is to be someone who would believe it if it were true, and who wouldn't believe it if it were false.

It will be useful to have a term for this situation when a person's belief is thus subjunctively connected to the fact. Let us say of a person who believes that *p*, which is true, that when (3) and (4) hold, his belief *tracks* the truth that *p*. To know is to have a belief that tracks the truth. Knowledge is a particular way of being connected to the world, having a specific real factual connection to the world: tracking it.

Scepticism

The sceptic about knowledge argues that we know very little or nothing of what we think we know, or at any rate that this position is no less reasonable than the belief in knowledge. The history of philosophy exhibits a number of different attempts to refute the sceptic: to prove him wrong or show that in arguing against knowledge he presupposes there is some and so refutes himself. Others attempt to show that accepting scepticism is unreasonable, since it is more likely that the sceptic's extreme conclusion is false than that all of his premisses are true, or simply because reasonableness of belief just means proceeding in an anti-sceptical way. Even when these counter-arguments satisfy their inventors, they fail to satisfy others, as is shown by the persistent attempts against scepticism. The continuing felt need to refute scepticism, and the difficulty in doing so, attests to the power of the sceptic's position, the depth of his worries.

An account of knowledge should illuminate sceptical arguments and show wherein lies their force. If the account leads us to reject these arguments, this had better not happen too easily or too glibly. To think the sceptic overlooks something obvious, to attribute to him a simple mistake or confusion or fallacy, is to refuse to acknowledge the power of his position and the grip it can have upon us. We thereby cheat ourselves of the opportunity to reap his insights and to gain self-knowledge in understanding why his arguments lure us so. Moreover, in fact, we cannot lay the spectre of scepticism to rest without first hearing what it shall unfold.

Our goal is not, however, to refute scepticism, to prove it is wrong or even to argue that it is wrong. We have elsewhere distinguished between philosophy that attempts to prove, and philosophy that attempts to explain how something is possible. Our task here is to explain how knowledge is possible, given what the sceptic says that we do accept (for example, that it is logically possible that we are dreaming or are floating in the tank). In doing this, we need not convince the sceptic, and we may introduce explanatory hypotheses that he would reject. What is important for our task of explanation and understanding is that we find those hypotheses acceptable or plausible, and that they show us how the existence of knowledge fits together with the logical possibilities the sceptic points to, so that these are reconciled within our own belief system. These hypotheses are to explain to ourselves how knowledge is possible, not to prove to someone else that knowledge is possible.[5,6]

Sceptical Possibilities

The sceptic often refers to possibilities in which a person would believe something even though it was false: really, the person is cleverly deceived by others, perhaps by an evil demon, or the person is dreaming, or he is floating in a tank near Alpha Centauri with his brain being stimulated. In each case, the *p* he believes is false, and he believes it even though it is false.

How do these possibilities adduced by the sceptic show that someone does not know that *p*? Suppose that someone is you; how do these possibilities count against your knowing that *p*? One way might be the following. (I shall consider other ways later.) If there is a possible situation where *p* is false yet you believe that *p*, then in that situation you believe that *p* even though it is false. So it appears you do not satisfy condition (3) for knowledge.

(3) If *p* were false, S wouldn't believe that *p*.

For a situation has been described in which you do believe that *p* even though *p* is false. How then can it also be true that if *p* were false, you wouldn't believe it? If the sceptic's possible situation shows that (3) is false, and if (3) is a necessary condition for knowledge, then the sceptic's possible situation shows that there isn't knowledge.

So construed, the sceptic's argument plays on condition (3); it aims to show that condition (3) is not satisfied. The sceptic may seem to be putting forth

[5]Gilbert Harman, *Thought* (Princeton; 1973), ch. 9, 142–54.

[6]From the perspective of explanation rather than proof, the extensive philosophical discussion, deriving from Charles S. Peirce, of whether the sceptic's doubts are real is beside the point. The problem of explaining how knowledge is possible would remain the same, even if no one ever claimed to doubt that there was knowledge.

R: Even if *p* were false, S still would believe *p*.

This conditional, with the same antecedent as (3) and the contradictory consequent, is incompatible with the truth of (3). If (3) is true, then R is not. However, R is stronger than the sceptic needs in order to show (3) is false. For (3) is false when if *p* were false, S might believe that *p*. This last conditional is weaker than R, and is merely (3)'s denial:

T: not-[not-*p* → not-(S believes that *p*)]

Whereas R does not simply deny (3), it asserts an opposing subjunctive of its own. Perhaps the possibility the sceptic adduces is not enough to show that R is true, but it appears at least to establish the weaker T; since this T denies (3), the sceptic's possibility appears to show that (3) is false. However, the truth of (3) is not incompatible with the existence of a possible situation where the person believes *p* though it is false. The subjunctive

(3) not-*p* → not-(S believes *p*)

does not talk of all possible situations in which *p* is false (in which not-*p* is true). It does not say that in all possible situations where not-*p* holds, S doesn't believe *p*. To say there is no possible situation in which not-*p* yet S believes *p*, would be to say that not-*p* entails not-(S believes *p*), or logically implies it. But subjunctive conditionals differ from entailments; the subjunctive (3) is not a statement of entailment. So the existence of a possible situation in which *p* is false yet S believes *p* does not show that (3) is false; (3) can be true even though there is a possible situation where not-*p* and S believes that *p*.

What the subjunctive (3) speaks of is the situation that would hold if *p* were false. Not every possible situation in which *p* is false is the situation that would hold if *p* were false. To fall into possible worlds talk, the subjunctive (3) speaks of the not-*p* world that is closest to the actual world, or of those not-*p* worlds that are closest to the actual world. And it is of this or these not-*p* worlds that it says (in them) S does not believe that *p*. What happens in yet other more distant not-*p* worlds is no concern of the subjunctive (3). The sceptic's possibilities (let us refer to them as SK), of the person's being deceived by a demon or dreaming or floating in a tank, count against the subjunctive

(3) if *p* were false then S wouldn't believe that *p*

only if (one of) these possibilities would or might obtain if *p* were false. Condition (3) says: if *p* were false, S still would not believe *p*. And this can hold even though there is some situation SK described by the sceptic in which *p* is false and S believes *p*. If *p* were false S still would not believe *p*, even though there is a situation SK in which *p* is false and S does believe *p*, provided that this situation SK wouldn't obtain if *p* were false. If the sceptic describes a situation SK which would not hold even if *p* were false then this situation SK doesn't show that (3) is false and so does not (in this way at least) undercut knowledge. Condition C acts to rule out sceptical hypotheses.

C: not-$p \rightarrow$ SK does not obtain.

Any sceptical situation SK which satisfies condition C is ruled out. For a sceptical situation SK to show that we don't know that p, it must fail to satisfy C which excludes it; instead it must be a situation that might obtain if p did not, and so satisfy C's denial:

not-(not-$p \rightarrow$ SK does not obtain)

Although the sceptic's imagined situations appear to show that (3) is false, they do not; they satisfy condition C and so are excluded.

The sceptic might go on to ask whether we know that his imagined situations SK are excluded by condition C, whether we know that if p were false SK would not obtain. However, typically he asks something stronger: do we know that his imagined situation SK does not actually obtain? Do we know that we are not being deceived by a demon, dreaming, or floating in a tank? And if we do not know this, how can we know that p? Thus we are led to the second way his imagined situations might show that we do not know that p.

Sceptical Results

According to our account of knowledge, S knows that the sceptic's situation SK doesn't hold if and only if

(1) SK doesn't hold
(2) S believes that SK doesn't hold
(3) If SK were to hold, S would not believe that SK doesn't hold
(4) If SK were not to hold, S would believe it does not.

Let us focus on the third of these conditions, The sceptic has carefully chosen his situations SK so that if they held we (still) would believe they did not. We would believe we weren't dreaming, weren't being deceived, and so on, even if we were. He has chosen situations SK such that if SK were to hold, S would (still) believe that SK doesn't hold – and this is incompatible with the truth of (3).

Since condition (3) is a necessary condition for knowledge, it follows that we do not know that SK doesn't hold. If it were true that an evil demon was deceiving us, if we were having a particular dream, if we were floating in a tank with our brains stimulated in a specified way, we would still believe we were not. So, we do not know we're not being deceived by an evil demon, we do not know we're not in that tank, and we do not know we're not having that dream. So says the sceptic, and so says our account. And also so we say – don't we? For how could we know we are not being deceived that way, dreaming that dream? If those things were happening to us, everything would seem the same to us. There is no way we can know it is not happening for there is no way we could tell if it were happening; and if it were happening we would believe exactly what we do now – in particular, we still would believe that it was not. For this reason, we feel, and correctly, that we don't know –

how could we? – that it is not happening to us. It is a virtue of our account that it yields, and explains, this result.

The sceptic asserts we do not know his possibilities don't obtain, and he is right. Attempts to avoid scepticism by claiming we do know these things are bound to fail. The sceptic's possibilities make us uneasy because, as we deeply realize, we do not know they don't obtain; it is not surprising that attempts to show we do know these things leave us suspicious, strike us even as bad faith. Nor has the sceptic merely pointed out something obvious and trivial. It comes as a surprise to realize that we do not know his possibilities don't obtain. It is startling, shocking. For we would have thought, before the sceptic got us to focus on it, that we did know those things, that we did know we were not being deceived by a demon, or dreaming that dream, or stimulated that way in that tank. The sceptic has pointed out that we do not know things we would have confidently said we knew. And if we don't know these things, what can we know? So much for the supposed obviousness of what the sceptic tells us.

Let us say that a situation (or world) is doxically identical for S to the actual situation when if S were in that situation, he would have exactly the beliefs (*doxa*) he actually does have. More generally, two situations are doxically identical for S if and only if he would have exactly the same beliefs in them. It might be merely a curiosity to be told there are non-actual situations doxically identical to the actual one. The sceptic, however, describes worlds doxically identical to the actual world in which almost everything believed is false.[7]

Such worlds are possible because we know mediately, not directly. This leaves room for a divergence between our beliefs and the truth. It is as though we possessed only two-dimensional plane projections of three-dimensional objects. Different three-dimensional objects, oriented appropriately, have the same two-dimensional plane projection. Similarly, different situations or worlds will lead to our having the very same beliefs. What is surprising is how very different the doxically identical world can be – different enough for almost everything believed in it to be false. Whether or not the mere fact that knowledge is mediated always makes room for such a very different doxically identical world, it does so in our case, as the sceptic's possibilities show. To be shown this is non-trivial, especially when we recall that we do not know the sceptic's possibility doesn't obtain: we do not know that we are not living in a doxically identical world wherein almost everything we believe is false.

What more could the sceptic ask for or hope to show? Even readers who sympathized with my desire not to dismiss the sceptic too quickly may feel this has gone too far, that we have not merely acknowledged the force of the sceptic's position but have succumbed to it.

[7] I say almost everything, because there still could be some true beliefs such as "I exist." More limited sceptical possibilities present worlds doxically identical to the actual world in which almost every belief of a' certain sort is false, for example, about the past, or about other people's mental states.

The sceptic maintains that we know almost none of what we think we know. He has shown, much to our initial surprise, that we do not know his (nontrivial) possibility SK doesn't obtain. Thus, he has shown of one thing we thought we knew, that we didn't and don't. To the conclusion that we know almost nothing, it appears but a short step. For if we do not know we are not dreaming or being deceived by a demon or floating in a tank, then how can I know, for example, that I am sitting before a page writing with a pen, and how can you know that you are reading a page of a book?

However, although our account of knowledge agrees with the sceptic in saying that we do not know that not-SK, it places no formidable barriers before my knowing that I am writing on a page with a pen. It is true that I am, I believe I am, if! weren't I wouldn't believe I was, and if I were, I would believe it. Also, it is true that you are reading a page (please, don't stop now!), you believe you are, if you weren't reading a page you wouldn't believe you were, and if you were reading a page you would believe you were. So according to the account, I do know that I am writing on a page with a pen, and you do know that you are reading a page. The account does not lead to any general scepticism.

Yet we must grant that it appears that if the sceptic is right that we don't know we are not dreaming or being deceived or floating in the tank, then it cannot be that I know I am writing with a pen or that you know you are reading a page. So we must scrutinize with special care the sceptic's "short step" to the conclusion that we don't know these things, for either this step cannot be taken or our account of knowledge is incoherent.

Nonclosure

In taking the "short step," the sceptic assumes that if S knows that p and he knows that "p entails q" then he also knows that q. In the terminology of the logicians, the sceptic assumes that knowledge is closed under known logical implication; that the operation of moving from something known to something else known to be entailed by it does not take us outside of the (closed) area of knowledge. He intends, of course, to work things backwards, arguing that since the person does not know that q, assuming (at least for the purposes of argument) that he does know that p entails q. it follows that he does not know that p. For if he did know that p, he would also know that q, which he doesn't.

The details of different sceptical arguments vary in their structure, but each one will assume some variant of the principle that knowledge is closed under known logical implication. If we abbreviate "knowledge that p" by "Kp" and abbreviate "entails" by the fishhook sign"$\hookrightarrow$," we can write this principle of closure as the subjunctive principle

$$p : \mathrm{K}(p \hookrightarrow q) \ \& \ \mathrm{K}p \ \rightarrow \ \mathrm{K}q.$$

If a person were to know that p entails q and he were to know that p then he would know that q. The statement that q follows by modus ponens from the other two stated as known in the antecedent of the subjunctive principle p; this principle counts on the person to draw the inference to q. You know that your being in a tank on Alpha Centauri entails your not being in place X where you are. (I assume here a limited readership.) And you know also the contrapositive, that your being at place X entails that you are not then in a tank on Alpha Centauri. If you knew you were at X you would know you're not in a tank (of a specified sort) at Alpha Centauri. But you do not know this last fact (the sceptic has argued and we have agreed) and so (he argues) you don't know the first. Another intuitive way of putting the sceptic's argument is as follows. If you know that two statements are incompatible and you know the first is true then you know the denial of the second.

You know that your being at X and your being in a tank on Alpha Centauri are incompatible; so if you knew you were at X you would know you were not in the (specified) tank on Alpha Centauri. Since you do not know the second, you don't know the first.

No doubt, it is possible to argue over the details of principle p, to point out it is incorrect as it stands. Perhaps, though Kp, the person does not know that he knows that p (that is, not-KKp) and so does not draw the inference to q. Or perhaps he doesn't draw the inference because not-KK($p \supset q$). Other similar principles face their own difficulties: for example, the principle that $K(p \rightarrow q) \rightarrow (Kp \rightarrow Kq)$ fails if Kp stops $p \rightarrow q$ from being true, that is, if $Kp \rightarrow$ not-$(p \rightarrow q)$; the principle that K($p \supset q$)- $\rightarrow$ K(K$p \rightarrow$ Kq) faces difficulties if Kp makes the person forget that ($p \supset q$) and so he fails to draw the inference to q. We seem forced to pile K upon K until we reach something like KK($p \supset q$) & KK$p \rightarrow$ Kq; this involves strengthening considerably the antecedent of p and so is not useful for the sceptic's argument that p is not known. (From a principle altered thus it would follow at best that it is not known that p is known.)

We would be ill-advised, however, to quibble over the details of p. Although these details are difficult to get straight, it will continue to appear that something like p is correct. If S knows that "p entails q," and he knows that p and knows that "(p and p entails q) entails q" and he does draw the inference to q from all this and believes q via the process of drawing this inference, then will he not know that q? And what is wrong with simplifying this mass of detail by writing merely principle p, provided we apply it only to cases where the mass of detail holds, as it surely does in the sceptical cases under consideration? For example, I do realize that my being in the Van Leer Foundation Building in Jerusalem entails that I am not in a tank on Alpha Centauri; I am capable of drawing inferences now; I do believe I am not in a tank on Alpha Centauri (though not solely via this inference, surely); and so forth. Won't this satisfy the correctly detailed principle, and shouldn't it follow that I know I am not (in that tank) on Alpha Centauri? The sceptic agrees it should follow; so he concludes from the fact that I don't know I am not floating in the tank on Alpha Centauri that I don't know I am in Jerusalem. Uncovering difficulties in the details of particular formulations of p will not weaken the principle's intuitive

appeal; such quibbling will seem at best like a wasp attacking a steamroller, at worst like an effort in bad faith to avoid being pulled along by the sceptic's argument.

Principle p is wrong, however, and not merely in detail. Knowledge is not closed under known logical implication. S knows that p when S has a true belief that p, and S wouldn't have a false belief that p (condition (3)) and S would have a true belief that p (condition (4)). Neither of these latter two conditions is closed under known logical implication. Let us begin with condition

(3) if p were false, S wouldn't believe that p.

When S knows that p, his belief that p is contingent on the truth of p. contingent in the way the subjunctive condition (3) describes. Now it might be that p entails q (and S knows this), that S's belief that p is subjunctively contingent on the truth of p, that S believes q, yet his belief that q is not subjunctively dependent on the truth of q. in that it (or he) does not satisfy:

(3′) if q were false, S wouldn't believe that q.

For (3′) talks of what S would believe if q were false, and this may be a very different situation from the one that would hold if p were false, even though p entails q. That you were born in a certain city entails that you were born on earth.[8] Yet contemplating what (actually) would be the situation if you were not born in that city is very different from contemplating what situation would hold if you weren't born on earth. Just as those possibilities are very different, so what is believed in them may be very different. When p entails q (and not the other way around) p will be a stronger statement than q, and so not-q (which is the antecedent of (3′)) will be a stronger statement than not-p (which is the antecedent of (3)). There is no reason to assume you will have the same beliefs in these two cases, under these suppositions of differing strengths.

There is no reason to assume the (closest) not-p world and the (closest) not-q world are doxically identical for you, and no reason to assume, even though p entails q, that your beliefs in one of these worlds would be a (proper) subset of your beliefs in the other.

Consider now the two statements:

p = I am awake and sitting on a chair in Jerusalem;
q = I am not floating in a tank on Alpha Centauri being stimulated by electrochemical means to believe that p.

The first one entails the second: p entails q. Also, I know that p entails q; and I know that p. If p were false, I would be standing or lying down in the same city, or perhaps sleeping there, or perhaps in a neighbouring city or town. If q were false, I would be floating in a tank on Alpha Centauri. Clearly these are very different situations, leading to great differences in what I then would believe. If p were false,

[8] Here again I assume a limited readership, and ignore possibilities such as those described in James Blish, *Cities in Flight* (New York, 1982).

if 1 weren't awake and sitting on a chair in Jerusalem, 1 would not believe that *p*. Yet if *q* were false, if I was floating in a tank on Alpha Centauri, 1 would believe that *q*, that I was not in the tank, and indeed, in that case, I would still believe that *p*. According to our account of knowledge, I know that *p* yet I do not know that *q*, even though (1 know) *p* entails *q*.

This failure of knowledge to be closed under known logical implication stems from the fact that condition (3) is not closed under known logical implication; condition (3) can hold of one statement believed while not of another known to be entailed by the first. It is clear that any account that includes as a necessary condition for knowledge the subjunctive condition (3), not-*p* → not-(S believes that *p*), will have the consequence that knowledge is not closed under known logical implication. When *p* entails *q* and you believe each of them, if you do not have a false belief that *p* (since *p* is true) then you do not have a false belief that *q*. However, if you are to know something not only don't you have a false belief about it, but also you wouldn't have a false belief about it. Yet, we have seen how it may be that *p* entails *q* and you believe each and you wouldn't have a false belief that *p* yet you might have a false belief that *q* (that is, it is not the case that you wouldn't have one). Knowledge is not closed under the known logical implication because 'wouldn't have a false belief that' is not closed under known logical implication.

If knowledge were the same as (simply) true belief then it would be closed under known logical implication (provided the implied statements were believed). Knowledge is not simply true belief, however; additional conditions are needed. These further conditions will make knowledge open under known logical implication, even when the entailed statement is believed, when at least one of the further conditions itself is open. Knowledge stays closed (only) if all of the additional conditions are closed. I lack a general non-trivial characterization of those conditions that are closed under known logical implication: possessing such an illuminating characterization, one might attempt to prove that no additional conditions of that sort could provide an adequate analysis of knowledge.

Still, we can say the following. A belief that *p* is knowledge that *p* only if it somehow varies with the truth of *p*. The causal condition for knowledge specified that the belief was "produced by" the fact, but that condition did not provide the right sort of varying with the fact. The subjunctive conditions (3) and (4) are our attempt to specify that varying. But however an account spells this out, it will hold that whether a belief that *p* is knowledge partly depends on what goes on with the belief in some situations when *p* is false. An account that says nothing about what is believed in any situation when *p* is false cannot give us any mode of varying with the fact.

Because what is preserved under logical implication is truth, any condition that is preserved under known logical implication is most likely to speak only of what happens when *p*, and *q*, are true, without speaking at all of what happens when either one is false. Such a condition is incapable of providing "varies with"; so adding only such conditions to true belief cannot yield an adequate account of knowledge.

A belief's somehow varying with the truth of what is believed is not closed under known logical implication. Since knowledge that *p* involves such variation,

knowledge also is not closed under known logical implication. The sceptic cannot easily deny that knowledge involves such variation, for his argument that we don't know that we're not floating in that tank, for example, uses the fact that knowledge does involve variation. ("If you were floating in the tank you would still think you weren't, so you don't know that you're not.") Yet, though one part of his argument uses that fact that knowledge involves such variation, another part of his argument presupposes that knowledge does not involve any such variation. This latter is the part that depends upon knowledge being closed under known logical implication. as when the sceptic argues that since you don't know that not-SK, you don't know you are not floating in the tank, then you also don't know, for example, that you are now reading a book. That closure can hold only if the variation does not. The sceptic cannot be right both times. According to our view he is right when he holds that knowledge involves such variation and so concludes that we don't know, for example, that we are not floating in that tank: but he is wrong when he assumes knowledge is closed under known logical implication and concludes that we know hardly anything.[9]

Knowledge is a real factual relation, subjunctively specifiable, whose structure admits our standing in this relation, tracking, to p without standing in it to some q which we know p to entail. Any relation embodying some variation of belief with this fact. with the truth (value), will exhibit this structural feature. The sceptic is right that we don't track Some particular truths – the ones stating that his sceptical possibilities SK don't hold – but wrong that we don't stand in the real knowledge-relation of tracking to many other truths, including ones that entail these first mentioned truths we believe but don't know.

The literature on scepticism contains writers who endorse these sceptical arguments (or similar narrower ones), but confess their inability to maintain their

[9]Reading an earlier draft of this chapter, friends pointed out to me that Fred Dretske already had defended the view that knowledge (as one among many epistemic concepts) is not closed under known logical implication. (See his "Epistemic Operators," *Journal of Philosophy*, 67, (1970). 1007–23.) Furthermore, Dretske presented a subjunctive condition for knowledge (in his "Conclusive Reasons," *Australasian Journal of Philosophy*, 49, (1971), 1–22), holding that S knows that p on the basis of reasons R only if: R would not be the case unless p were the case. Here Dretske ties the evidence subjunctively to the fact, and the belief based on the evidence subjunctively to the fact through the evidence. The independent statement and delineation of the position here I hope will make clear its many merits.

After Goldman's paper on a causal theory of knowledge, in *Journal of Philosophy*, 64, (1967), an idea then already "in the air," it required no great leap to consider subjunctive conditions. Some 2 months after the first version of this chapter was written, Goldman himself published a paper on knowledge utilizing counterfactuals ("Discrimination and Perceptual Knowledge," Essay II in this collection), also talking of relevant possibilities (without using the counterfactuals to identify which possibilities are relevant); and R. Shope has called my attention to a paper of L. S. Carrier ("An Analysis of Empirical Knowledge," *Southern Journal of Philosophy*, 9, (1971), 3–11) that also used subjunctive conditions including our condition (3). Armstrong's reliability view of knowledge (Belief, Truth and Knowledge, Cambridge, 1973, pp. 166, 169) involved a lawlike connection between the belief that p and the state of affairs that makes it true. Clearly, the idea is one whose time has come.

sceptical beliefs at times when they are not focusing explicitly on the reasoning that led them to sceptical conclusions. The most notable example of this is Hume:

> I am ready to reject all belief and reasoning, and can look upon no opinion even as more probable or likely than another. . . Most fortunately it happens that since reason is incapable of dispelling these clouds, nature herself suffices to that purpose, and cures me of this philosophical melancholy and delirium, either by relaxing this bent of mind, or by some avocation, and lively impression of my senses, which obliterate all these chimeras. I dine, I play a game of backgammon, I converse, and am merry with my friends; and when after three or four hours' amusement, I would return to these speculations, they appear so cold, and strained, and ridiculous, that I cannot find in my heart to enter into them any farther. (A Treatise of Human Nature, Book I, Part IV, section VII.)
>
> The great subverter of Pyrrhonism or the excessive principles of skepticism is action, and employment, and the occupations of common life. These principles may flourish and triumph in the schools; where it is, indeed, difficult, if not impossible, to refute them. But as soon as they leave the shade, and by the presence of the real objects, which actuate our passions and sentiments, are put in opposition to the more powerful principles of our nature, they vanish like smoke, and leave the most determined skeptic in the same condition as other mortals. . . And though a Pyrrhonian may throw himself or others into a momentary amazement and confusion by his profound reasonings; the first and most trivial event in life will put to flight all his doubts and scruples, and leave him the same, in every point of action and speculation, with the philosophers of every other sect, or with those who never concerned themselves in any philosophical researches. When he awakes from his dream, he will be the first to join in the laugh against himself, and to confess that all his objections are mere amusement. (An Enquiry Concerning Human Understanding, Section XII, Part II.)

The theory of knowledge we have presented explains why sceptics of various sorts have had such difficulties in sticking to their far-reaching sceptical conclusions "outside the study," or even inside it when they are not thinking specifically about sceptical arguments and possibilities SK.

The sceptic's arguments do show (but show only) that we don't know the sceptic's possibilities SK do not hold; and he is right that we don't track the fact that SK does not hold. (If it were to hold, we would still think it didn't.) However, the sceptic's arguments don't show we do not know other facts (including facts that entail not-SK) for we do track these other facts (and knowledge is not closed under known logical entailment). Since we do track these other facts – you, for example, the fact that you are reading a book; I, the fact that I am writing on a page – and the sceptic tracks such facts too, it is not surprising that when he focuses on them, on his relationship to such facts, the sceptic finds it hard to remember or maintain his view that he does not know those facts. Only by shifting his attention back to his relationship to the (different) fact that not-SK, which relationship is not tracking, can he revive his sceptical belief and make it salient. However, this sceptical triumph is evanescent, it vanishes when his attention turns to other facts. Only by fixating on the sceptical possibilities SK can he maintain his sceptical virtue; otherwise, unsurprisingly, he is forced to confess to sins of credulity.

Chapter 30
On Logics of Knowledge and Belief

Robert Stalnaker

Introduction

Formal epistemology, or at least the approach to formal epistemology that develops a logic and formal semantics of knowledge and belief in the possible worlds framework, began with Jaakko Hintikka's book *Knowledge and Belief*, published in 1962. Hintikka's project sparked some discussion of issues about iterated knowledge (does knowing imply knowing that one knows?) and about "knowing who", and quantifying into knowledge attributions. Much later, this kind of theory was taken up and applied by theoretical computer scientists and game theorists.[1] The formal semantic project gained new interest when it was seen that it could be applied to contexts with multiple knowers, and used to clarify the relation between epistemic and other modal concepts.

Edmund Gettier's classic refutation of the Justified True Belief analysis of knowledge (Gettier 1963) was published at about the same time as Hintikka's book, and it immediately spawned an epistemological industry—a project of attempting to revise the refuted analysis by adding further conditions to meet the counterexamples. Revised analyses were met with further counterexamples, followed by further refinements. This kind of project flourished for some years, but eventually became an internally driven game that was thought to have lost contact with the fundamental epistemological questions that originally motivated it. This way of approaching epistemological questions now seems hopelessly out of date,

[1] See Fagin et al. (1995) and Battigalli and Bonanno (1999) for excellent surveys of the application of logics of knowledge and belief in theoretical computer science and game theory.

R. Stalnaker (✉)
Department of Linguistics and Philosophy, MIT, Cambridge, MA, USA
e-mail: stal@mit.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_30

but I think there may still be some insights to be gained by looking back, if not at the details of the analyses, at some of the general strategies of analysis that were deployed.

There was little contact between these two very different epistemological projects. The first had little to say about substantive questions about the relation between knowledge, belief, and justification or epistemic entitlement, or about traditional epistemological issues, such as skepticism. The second project ignored questions about the abstract structure of epistemic and doxastic states. But I think some of the abstract questions about the logic of knowledge connect with traditional questions in epistemology, and with the issues that motivated the attempt to find a definition of knowledge. The formal semantic framework provides the resources to construct models that may help to clarify the abstract relationship between the concept of knowledge and some of the other concepts (belief and belief revision, causation and counterfactuals) that were involved in the post-Gettier project of defining knowledge. And some of the examples that were originally used in the post-Gettier literature to refute a proposed analysis can be used in a different way in the context of formal semantic theories: to bring out contrasting features of some alternative conceptions of knowledge, conceptions that may not provide plausible analyses of knowledge generally, but that may provide interesting models of knowledge that are appropriate for particular applications, and that may illuminate, in an idealized way, one or another of the dimensions of the complex epistemological terrain.

My aim in this paper will be to bring out some of the connections between issues that arise in the development and application of formal semantics for knowledge and belief and more traditional substantive issues in epistemology. The paper will be programmatic; pointing to some highly idealized theoretical models, some alternative assumptions that might be made about the logic and semantics of knowledge, and some of the ways in which they might connect with traditional issues in epistemology, and with applications of the concept of knowledge. I will bring together and review some old results, and make some suggestions about possible future developments. After a brief sketch of Hintikka's basic logic of knowledge, I will discuss, in section "Partition models", the S5 epistemic models that were developed and applied by theoretical computer scientists and game theorists, models that, I will argue, conflate knowledge and belief. In section "Belief and knowledge", I will discuss a basic theory that distinguishes knowledge from belief and that remains relatively noncommittal about substantive questions about knowledge, but that provides a definition of belief in terms of knowledge. This theory validates a logic of knowledge, S4.2, that is stronger than S4, but weaker than S5. In the remaining four sections, I will consider some alternative ways of adding constraints on the relation between knowledge and belief that go beyond the basic theory: in section "Partition models and the basic theory", I will consider the S5 partition models as a special case of the basic theory; in section "Minimal and maximal extensions", I will discuss the upper and lower bounds to an extension of the semantics of belief to a semantics for knowledge; in section "Belief revision and the defeasibility analysis",

I will discuss a version of the defeasibility analysis of knowledge, and in section "The causal dimension", a simplified version of a causal theory.

The basic idea that Hintikka developed, and that has since become familiar, was to treat knowledge as a modal operator with a semantics that parallels the possible worlds semantics for necessity. Just as necessity is truth in all possible worlds, so knowledge is truth in all *epistemically* possible worlds. The assumption is that to have knowledge is to have a capacity to locate the actual world in logical space, to exclude certain possibilities from the candidates for actuality. The epistemic possibilities are those that remain after the exclusion, those that the knower cannot distinguish from actuality. To represent knowledge in this way is of course not to provide any kind of reductive analysis of knowledge, since the abstract theory gives no substantive account of the criteria for determining epistemic possibility. The epistemic possibilities are defined by a binary accessibility relation between possible worlds that is a primitive component of an epistemic model. (Where x and y are possible worlds, and 'R' is the accessibility relation, 'xRy' says that y is epistemically possible for the agent in world x.). The idea was to give a precise representation of the structure of an epistemic state that was more or less neutral about more substantive questions about what constitutes knowledge, but that sharpened questions about the logic of knowledge. This form of representation was, however, far from innocent, since it required, from the start, an extreme idealization: Even in its most neutral form, the framework required the assumption that knowers know all logical truths, and all of the consequences of their knowledge, since no matter how the epistemically possible worlds are selected, all logical truths will be true in all of them, and for any set of propositions true in all of them, all of their logical consequences will also be true in all of them. There are different ways of understanding the character of this idealization: on the one hand, one might say that the concept of knowledge that is being modeling is knowledge in the ordinary sense, but that the theory is intended to apply only to idealized knowers—those with superhuman logical capacities. Alternatively, one might say that the theory is intended to model an idealized sense of knowledge—the information that is implicit in one's knowledge—that literally applies to ordinary knowers. However the idealization is explained, there remain the questions whether it is fruitful to develop a theory that requires this kind of deviation from reality, and if so why.[2] But I think these questions are best answered by looking at the details of the way such theories have been, and can be developed.

The most basic task in developing a semantics for knowledge in the possible worlds framework is to decide on the properties of the epistemic accessibility relation. It is clear that the relation should be reflexive, which is necessary to validate the principle that knowledge implies truth, an assumption that is just about the only principle of a logic of knowledge that is uncontroversial. Hintikka argued

[2]I explore the problem of logical omniscience in two papers, Stalnaker (1991, 1999b) both included in Stalnaker (1999a). I don't attempt to solve the problem in either paper, but only to clarify it, and to argue that it is a genuine problem, and not an artifact of a particular theoretical framework.

that we should also assume that the relation is transitive, validating the much more controversial principle that knowing implies knowing that one knows. Knowing and knowing that one knows are, Hintikka claimed, "virtually equivalent." Hintikka's reasons for this conclusion were not completely clear. He did not want to base it on a capacity for introspection: he emphasized that his reasons were logical rather than psychological. His proof of the KK principle rests on the following principle: If $\{K\varphi, \sim K\sim\psi\}$ is consistent, then $\{K\varphi, \psi\}$ is consistent, and it is clear that if one grants this principle, the KK principle immediately follows.[3] The reason for accepting this principle seems to be something like this: Knowledge requires conclusive reasons for belief, reasons that would not be defeated by any information compatible with what is known. So if one knows that φ while ψ is compatible with what one knows, then the truth of ψ could not defeat one's claim to know that φ. This argument, and other considerations for and against the KK principle deserve more careful scrutiny. There is a tangle of important and interesting issues underlying the question whether one should accept the KK principle, and the corresponding semantics, and some challenging arguments that need to be answered if one does.[4] I think the principle can be defended (in the context of the idealizations we are making), but I will not address this issue here, provisionally following Hintikka in accepting the KK principle, and a semantics that validates it.

The S4 principles (Knowledge implies truth, and knowing implies knowing that one knows) were as far as Hintikka was willing to go. He unequivocally rejects the characteristic S5 principle that if one lacks knowledge, then one knows that one lacks it.("unless you happen to be as sagacious as Socrates"[5]), and here his reasons seem to be clear and decisive[6]:

> The consequences of this principle, however, are obviously wrong. By its means (together with certain intuitively acceptable principles) we could, for example, show that the following sentence is self sustaining:
>
> (13) $p \supset K_a P_a p$. [In Hintikka's notation, 'P_a' is the dual of the knowledge operator, 'K_a': '$\sim K_a \sim$'. I will use 'M' for $\sim$K$\sim$]

The reason that (13) is clearly unacceptable, as Hintikka goes on to say, is that it implies that one could come to know by reflection alone, of any truth, that it was compatible with one's knowledge. But it seems that a consistent knower might believe, and be justified in believing, that she knew something that was in fact false. That is, it might be, for some proposition φ that $\sim\varphi$, and $BK\varphi$. In such a case, if the subject's beliefs are consistent, then she does not believe, and so does not know, that $\sim\varphi$ is compatible with her knowledge. That is, $\sim K\sim K\varphi$, along with $\sim\varphi$, will be true, falsifying (13).

[3] Substituting '$\sim K\varphi$' for ψ, and eliminating a double negation, the principle says that if $\{K\varphi, \sim KK\varphi\}$ is consistent, then $\{K\varphi, \sim K\varphi\}$ is consistent.

[4] See especially, Williamson (2000) for some reasons to reject the KK principle.

[5] Hintikka (1962, 106).

[6] Ibid, 54.

Partition Models

Despite Hintikka's apparently decisive argument against the S5 principle, later theorists applying epistemic logic and semantics, both in theories of distributive computer systems and in game theory assumed that S5 was the right logic for (an idealized concept of) knowledge, and they developed semantic models that seem to support that decision. But while such models, properly interpreted, have their place, I will argue that they have conflated knowledge and belief in a way that has led to some conceptual confusion, and that they have abstracted away from some interesting problems within their intended domains of application that more general models might help to clarify. But before getting to this issue, let me first take note of another way that more recent theorists have modified, or generalized, Hintikka's original theory.

Hintikka's early models were models of the knowledge of a single knower, but much of the later interest in formal epistemic models derives from a concern with situations in which there are multiple knowers who may know or be ignorant about the knowledge and ignorance of the others. While Hintikka's early work did not give explicit attention to the interaction of different knowers, the potential to do so is implicit in his theory. Both the logic and the semantics of the knowledge of a single knower generalize in a straightforward way to a model for multiple knowers. One needs only a separate knowledge operator for each knower, and in the semantics, a separate relation of epistemic accessibility for each knower that interprets the operator. One can also introduce, for any group of knowers, an operator for the *common* knowledge shared by the member of the group, where a group has common knowledge that φ if and only if all know that φ, all know that all know that φ, all know that all know that all know, etc. all the way up. The semantics for the common knowledge operator is interpreted in terms of an accessibility relation that is definable in terms of the accessibility relations for the individual knowers: the common-knowledge accessibility relation for a group G is the transitive closure of the set of epistemic accessibility relations for the members of that group.[7] If R^G is this relation, then the knowers who are members of G have common knowledge that φ (in possible world x) iff φ is true in all possible worlds that are R^G related to world x. The generalization to multiple knowers and to common knowledge, works the same way, whatever assumptions one makes about the accessibility relation, and one can define notions of common belief in an exactly analogous way. The properties of the accessibility relations for common knowledge and common belief will derive from the properties of the individual accessibility relations, but they won't necessarily be the same as the properties of the individual

[7]More precisely, if R^i is the accessibility relation for knower i, then the common-knowledge accessibility relation for a group G is defined as follows; xR^Gy iff there is a sequence of worlds, $z_1, \ldots z_n$ such that $z_1 = x$ and $z_n = y$ and for all j between 1 and n−1, there is a knower $i \in G$, such that $z_j R^i z_{j+1}$.

accessibility relations. (Though if the logic of knowledge is S4 or S5, then the logic of common knowledge will also be S4 or S5, respectively).

Theoretical computer scientists have used the logic and semantics for knowledge to give abstract descriptions of distributed computer systems (such as office networks or email systems) that represent the distribution and flow of information among the components of the system. For the purpose of understanding how such systems work and how to design protocols that permit them to accomplish the purposes for which they are designed, it is useful to think of them as communities of interacting rational agents who use what information they have about the system as a whole to serve their own interests, or to play their part in a joint project. And it is useful in turn for those interested in understanding the epistemic states of rational agents to think of them in terms of the kind of simplified models that theoretical computer scientists have constructed.

A distributed system consists of a set of interconnected components, each capable of being in a range of local states. The way the components are connected, and the rules by which the whole system works, constrain the configurations of states of the individual components that are possible. One might specify such a system by positing a set of n components and possible local states for each. One might also include a component labeled "nature" whose local states represent information from outside the system proper. *Global* states will be n-tuples of local states, one for each component, and the model will also specify the set of global states that are *admissible*. Admissible global states are those that are compatible with the rules governing the way the components of the system interact. The admissible global states are the possible worlds of the model. This kind of specification will determine, for each local state that any component might be in, a set of global states (possible worlds) that are compatible with the component being in that local state. This set will be the set of epistemically possible worlds that determines what the component in that state knows about the system as a whole.[8] Specifically, if 'a' and 'b' denote admissible global states, and 'a_i' and 'b_i' denote the ith elements of a and b, respectively (the local states of component i.), then global world-state b is epistemically accessible (for i) to global world-state b if and only if $a_i = b_i$. So, applying the standard semantic rule for the knowledge operator, component (or knower) i will know that φ, in possible world a, if and only if φ is true

[8] A more complex kind of model would specify a set of admissible *initial* global states, and a set of transition rules taking global states to global states. The possible worlds in this kind of model are the admissible global *histories*—the possible ways that the system might evolve. In this kind of model, one can represent the distribution of information, not only about the current state of the system, but also about how it evolved, and where it is going. In the more general model, knowledge states are time-dependent, and the components may have or lack information not only about which possible world is actual, but also about where (temporally) it is in a given world. The dynamic dimension, and the parallels with issues about indexical knowledge and belief, are part of the interest of the distributed systems models, but I will ignore these issues here.

in all possible worlds in which i has the same local state that it has in world-state a. One knows that φ if one's local state carries the information that φ.[9]

Now it is obvious that this epistemic accessibility relation is an equivalence relation, and so the logic for knowledge in a model of this kind is S5. Each of the epistemic accessibility relations partitions the space of possible worlds, and the cross-cutting partitions give rise to a simple and elegant model of common knowledge, also with an S5 logic. Game theorists independently developed this kind of partition model of knowledge and have used such models to bring out the consequences of assumptions about common knowledge. For example, it can be shown that, in certain games, players will always make certain strategy choices when they have common knowledge that all players are rational. But as we have seen, Hintikka gave reasons for rejecting the S5 logic for knowledge, and the reasons seemed to be decisive. It seems clear that a consistent and epistemically responsible agent might take herself to know that φ in a situation in which φ was in fact false. Because knowledge implies truth, it would be false, in such a case, that the agent knew that φ, but the agent could not know that she did not know that φ without having inconsistent beliefs. If such a case is possible, then there will be counterexamples to the S5 principle, $\sim K\varphi \rightarrow K \sim K\varphi$. That is, the S5 principles require that rational agents be immune to error. It is hard to see how any theory that abstracts away from the possibility of error could be relevant to epistemology, an enterprise that begins with skeptical arguments using scenarios in which agents are systematically mistaken and that seeks to explain the relation between knowledge and belief, presupposing that these notions do not coincide.

Different theorists have different purposes, and it is not immediately obvious that the models of knowledge that are appropriate to the concerns of theoretical computer scientists and game theorists need be relevant to issues in epistemology. But I think that the possibility of error, and the differences between knowledge and belief are relevant to the intended domains of application of those models, and that some of the puzzles and problems that characterize epistemology are reflected in problems that may arise in applying those theories.

As we all know too well, computer systems sometimes break down or fail to behave as they were designed to behave. In such cases, the components of a distributed system will be subject to something analogous to error and illusion. Just as the epistemologist wants to explain how and when an agent knows some things even when he is in error about others, and is interested in methods of detecting and avoiding error, so the theoretical computer scientist is interested in the way that the components of a system can avoid and detect faults, and can continue to

[9]Possible worlds, on this way of formulating the theory, are not primitive points, as they are in the usual abstract semantics, but complex objects—sequences of local states. But an equivalent formulation might begin with a given set of primitive (global) states, together with a set of equivalence relations, one for each knower, and one for "nature". The local states could then be defined as the equivalence classes.

function appropriately even when conditions are not completely normal. To clarify such problems, it is useful to distinguish knowledge from something like belief.

The game theorist, or any theorist concerned with rational action, has a special reason to take account of the possibility of false belief, even under the idealizing assumption that in the actual course of events, everyone's beliefs are correct. The reason is that decision theorists and game theorists need to be concerned with causal or counterfactual possibilities, and to distinguish them from epistemic possibilities. When I deliberate, or when I reason about why it is rational to do what I know that I am going to do, I need to consider possible situations in which I make alternative choices. I know, for example, that it would be irrational to cooperate in a one-shot prisoners' dilemma because I know that in the counterfactual situation in which I cooperate, my payoff is less than it would be if I defected. And while I have the capacity to influence my payoff (negatively) by making this alternative choice, I could not, by making this choice, influence your prior beliefs about what I will do; that is, your prior beliefs will be the same, in the counterfactual situation in which I make the alternative choice, as they are in the actual situation Since you take yourself (correctly, in the actual situation) to know that I am rational, and so that I will not cooperate, you therefore also take yourself to know, in the counterfactual situation I am considering, that I am rational, and so will not cooperate. But in that counterfactual situation, you are wrong—you have a false belief that you take to be knowledge. There has been a certain amount of confusion in the literature about the relation between counterfactual and epistemic possibilities, and this confusion is fed, in part, by a failure to make room in the theory for false belief.[10]

Even in a context in which one abstracts away from error, it is important to be clear about the nature of the idealization, and there are different ways of understanding it that are sometimes confused. But before considering the alternative ways of making the S5 idealization, let me develop the contrast between knowledge and belief, and the relation between them, in a more general setting.

Belief and Knowledge

Set aside the S5 partition models for the moment, and consider, from a more neutral perspective, the logical properties of belief, and the relation between belief and knowledge. It seems reasonable to assume, at least in the kind of idealized context we are in, that agents have introspective access to their beliefs: if they believe that φ, then they know that they do, and if they do not, then they know that they do not. (The S5, "negative introspection" principle, $\sim K\varphi \rightarrow K \sim K\varphi$, was problematic for knowledge because it is in tension with the fact that knowledge implies truth, but the corresponding principle for belief does not face this problem.) It also seems reasonable to assume that knowledge implies belief. Given the fact that our idealized

[10]These issues are discussed in Stalnaker (1996).

believers are logically omniscient, we can assume, in addition, that their beliefs will be consistent. Finally, to capture the fact that our intended concept of belief is a strong one—subjective certainty—we assume that believing implies believing that one knows. So our logic of knowledge and belief should include the following principles in addition to those of the logic S4:

(PI)	$\vdash B\varphi \rightarrow KB\varphi$	positive introspection
(NI)	$\vdash \sim B\varphi \rightarrow K \sim B\varphi$	negative introspection
(KB)	$\vdash K\varphi \rightarrow B\varphi$	knowledge implies belief
(CB)	$\vdash B\varphi \rightarrow \sim B \sim \varphi$	consistency of belief
(SB)	$\vdash B\varphi \rightarrow BK\varphi$	strong belief

The resulting combined logic for knowledge and belief yields a pure belief logic, KD45, which is validated by a doxastic accessibility relation that is serial, transitive and euclidean.[11] More interestingly, one can prove the following equivalence theorem: $\vdash B\varphi \leftrightarrow MK\varphi$ (using 'M' as the epistemic possibility operator, '$\sim$K$\sim$'). This equivalence permits a more economical formulation of the combined belief-knowledge logic in which the belief operator is defined in terms of the knowledge operator. If we substitute 'MK' for 'B' in our principle (CB), we get $MK\varphi \rightarrow KM\varphi$, which, if added to S4 yields the logic of knowledge, S4.2. All of the other principles listed above (with 'MK' substituted for 'B') are theorems of S4.2, so this logic of knowledge by itself yields a combined logic of knowledge and belief with the appropriate properties.[12]

The assumptions that are sufficient to show the equivalence of belief with the epistemic possibility of knowledge (one believes that φ, in the strong sense, if and only if it is compatible with one's knowledge that one knows that φ) might also be made for a concept of *justified* belief, although the corresponding assumptions will be more controversial. Suppose (1) one assumes that justified belief is a necessary condition for knowledge, and (2) one adopts an *internalist* conception of justification that supports the positive and negative introspection conditions (if one has justified belief that φ, one knows that one does, and if one does not, one knows that one does not), and (3) one assumes that since the relevant concept of belief is a strong one, one is justified in believing that φ if and only if one is justified in believing that one knows that φ. Given these assumptions, justified belief will also

[11] KD45 adds to the basic modal system K the axioms (D), which is our CB, (4) $B\varphi \rightarrow BB\varphi$, which follows immediately from our (PI) and (KB), and (5) $\sim B\varphi \rightarrow B \sim B\varphi$, which follows immediately from (NI) and (KB). The necessitation rule for B (If $\vdash \varphi$, then $\vdash B\varphi$) and the distribution principle ($B(\varphi \rightarrow \psi) \rightarrow (B\varphi \rightarrow B\psi)$) can both be derived from our principles.

[12] The definability of belief in terms of knowledge, and the point that the assumptions about the relation between knowledge and belief imply that the logic of knowledge should be S4.2, rather than S4, were first shown by Wolfgang Lenzen. See his classic monograph, *Recent Work in Epistemic Logic. Acta Philosophica Fennica* **30** (1978). North-Holland, Amsterdam.

coincide with the epistemic possibility that one knows, and so belief and justified belief will coincide. The upshot is that for an internalist, a divergence between belief (in the strong sense) and justified belief would be a kind of internal inconsistency. If one is not fully justified in believing φ, one knows this, and so one knows that a necessary condition for knowledge that φ is lacking. But if one believes that φ, in the strong sense, then one believes that one knows it. So one both knows that one lacks knowledge that φ, and believes that one has knowledge that φ.

The usual constraint on the accessibility relation that validates S4.2 is the following convergence principle (added to the transitivity and reflexivity conditions): if xRy and xRz, then there is a w such that yRw and zRw. But S4.2 is also sound and complete relative to the following stronger convergence principle: for all x, there is a y such that for all z, if xRz, then zRy. The weak convergence principle (added to reflexivity and transitivity) implies that for any *finite* set of worlds accessible to x, there is a single world accessible with respect to all of them. The strong convergence principle implies that there is a world that is accessible to *all* worlds that are accessible to x. The semantics for our logic of knowledge requires the stronger convergence principle.[13]

Just as, within the logic, one can define belief in terms of knowledge, so within the semantics, one can define a doxastic accessibility relation for the derived belief operator in terms of the epistemic accessibility relation. If 'R' denotes the epistemic accessibility relation and 'D denotes the doxastic relation, then the definition is as follows: $xDy =_{df} (z)\ (xRz \rightarrow zRy)$. Assuming that R is transitive, reflexive and strongly convergent, it can be shown that D will be serial, transitive and euclidean—the constraints on the accessibility relation that characterize the logic KD45.

One can also define, in terms of D, and so in terms of R, a third binary relation on possible worlds that is relevant to describing the epistemic situation of our ideal knower: Say that two possible worlds x and y are *epistemically indistinguishable* to an agent (xEy) if and only if she has exactly the same beliefs in world x as she has in world y. That is, $xEy =_{df} (z)\ (xDz \leftrightarrow yDz)$. E is obviously an equivalence relation, and so any modal operator interpreted in the usual way in terms of E would be an S5 operator But while this relation is definable in the semantics in terms of the epistemic accessibility relation, we cannot define, in the object language with just the knowledge operator, a modal operator whose semantics is given by this accessibility relation.

So the picture that our semantic theory paints is something like this: For any given knower i and possible world x, there is, first, a set of possible worlds that are subjectively indistinguishable from x, to i (those worlds that are E-related to x); second, there is a subset of that set that includes just the possible worlds compatible with what i *knows* in x (those worlds that are R-related to x); third, there is a subset of

[13]The difference between strong and weak convergence does not affect the *propositional* modal logic, but it will make a difference to the quantified modal logic. The following is an example of a sentence that is valid in models satisfying strong convergence (along with transitivity and reflexivity) but not valid in all models satisfying weak convergence: $M K ((x)\ (MK\varphi \rightarrow \varphi))$.

that set that includes just the possible worlds that are compatible with what i *believes* in x (those worlds that are D-related to x). The world x itself will necessarily be a member of the outer set and of the R-subset, but will not necessarily be a member of the inner D-subset. But if x is itself a member of the inner D-set (if world x is itself compatible with what i believes in x), then the D-set will coincide with the R-set.

Here is one way of seeing this more general theory as a generalization of the distributive systems models, in which possible world-states are sequences of local states: one might allow *all* sequences of local states (one for each agent) to count as possible world-states, but specify, for each agent, a subset of them that are *normal*—the set in which the way that agent interacts with the system as a whole conforms to the constraints that the system conforms to when it is functioning as it is supposed to function. In such models, two worlds, x and y, will be subjectively indistinguishable, for agent i ($xE_i\ y$), whenever $x_i = y_i$ (so the relation that was the epistemic accessibility relation in the unreconstructed S5 distributed systems model is the subjective indistinguishability relation in the more general models). Two worlds are related by the doxastic accessibility relation ($xD_i\ y$) if and only if $x_i = y_i$, *and in addition, y is a normal world, with respect to agent i*.[14] This will impose the right structure on the D and E relations, and while it imposes some constraints on the epistemic accessibility relation, it leaves it underdetermined. We might ask whether R can be defined in a plausible way in terms of the components of the model we have specified, or whether one might add some independently motivated components to the definition of a model that would permit an appropriate definition of R. This question is a kind of analogue of the question asked in the more traditional epistemological enterprise—the project of giving a definition of knowledge in terms of belief, truth, justification, and whatever other normative and causal concepts might be thought to be relevant. Transposed into the model theoretic framework, the traditional problem of adding to true belief further conditions that together are necessary and sufficient for knowledge is the problem of extending the doxastic accessibility relation to a reflexive relation that is the right relation (at least in the idealized context) for the interpretation of a knowledge operator. In the remainder of this paper, I will consider several ways that this might be done, and at the logics of knowledge that they validate.

[14]We observed in note 7 that an equivalent formulation of the S5 distributed systems models would take the global world-states as primitive, specifying an equivalence relation for each agent, and defining local states as equivalence classes of global states. In an equivalent formulation of this kind of the more general theory, the assumption that every sequence of local states is a possible world will be expressed by a recombination condition: that for every sequence of equivalence classes (one for each agent) there is a possible world that is a member of their intersection. I have suggested that a recombination condition of this kind should be imposed on game theoretic models (where the equivalence classes are types, represented by probability functions), defending it as a representation of the conceptual independence of the belief states of different agents.

Partition Models and the Basic Theory

One extreme way of defining the epistemic accessibility relation in terms of the resources of our models is to identify it with the relation of subjective indistinguishability, and this is one way that the S5 partition models have implicitly been interpreted. If one simply assumes that the epistemic accessibility relation is an equivalence relation, this will suffice for a collapse of our three relations into one. Subjective indistinguishability, knowledge, and belief will all coincide. This move imposes a substantive condition on knowledge, and so on belief, when it is understood in the strong sense as belief that one knows, a condition that is appropriate for the skeptic who thinks that we are in a position to have genuine knowledge only about our own internal states—states about which we cannot coherently be mistaken. On this conception of knowledge, one can have a false belief (in the strong sense) only if one is internally inconsistent, and so this conception implies a bullet-biting response to the kind of argument that Hintikka gave against the S5 logic for knowledge. Hintikka's argument was roughly this: S5 validates the principle that any proposition that is in fact false, is known by any agent to be compatible with his knowledge, and this is obviously wrong: The response suggested by the conception of knowledge that identifies knowledge with subjective indistinguishability is that if we assume that all we can know is how things seem to us, and also assume that we are infallible judges of the way things seem to us, then it will be reasonable to conclude that we are in a position to know, of anything that is in fact false, that we do not know it.

There is a less radical way to reconcile our basic theory of knowledge and belief with the S5 logic and the partition models. Rather than making more restrictive assumptions about the concept of knowledge, or about the basic structure of the model, one may simply restrict the intended domain of application of the theory to cases in which the agent in question has, in fact, only true beliefs. On this way of understanding the S5 models, the model theory does not further restrict the relations between the three accessibility relations, but instead assumes that the *actual* world of the model is a member of the inner D-set.[15] This move does not provide us with

[15]In most formulations of a possible-worlds semantics for propositional modal logic, a *frame* consists simply of a set of worlds and an accessibility relation. A model on a frame determines the truth values of sentences, relative to each possible world. On this conception of a model, one cannot talk of the truth of a sentence in a model, but only of truth *at a world* in a model. Sentence validity is defined, in formulations of this kind, as truth in all worlds in all models. But in some formulations, including in Kripke's original formal work, a frame (or model structure, as Kripke called it at the time) included, in addition to a set of possible worlds and an accessibility relation, a designated possible world—the actual world of the model. A sentence is true in a model if it is true in the designated actual world, and valid if true in all models. This difference in formulation was a minor detail in semantic theories for most of the normal modal logics, since any possible world of a model might be the designated actual world without changing anything else. So the two ways of defining sentence validity will coincide. But the finer-grained definition of a frame allows for theories in which the constraints on R, and the semantic rules for operators, make reference to the

a way to define the epistemic accessibility relation in terms of the other resources of the model; but what it does is to stipulate that the actual world of the model is one for which the epistemic accessibility relation is determined by the other components. (That is, the set of worlds y that are epistemically accessible to the actual world is determined) Since the assumptions of the general theory imply that all worlds outside the D-sets are epistemically inaccessible to worlds within the D-sets, and that all worlds within a given D-set are epistemically accessible to each other, the assumption that the actual world of the model is in a D-set will determine the R-set for the actual world, and will validate the logic S5.

So long as the object language that is being interpreted contains just one modal operator, an operator representing the knowledge of a single agent, the underdetermination of epistemic accessibility will not be reflected in the truth values in a model of any expressible proposition. Since all possible worlds outside of any D-set will be invisible to worlds within it, one could drop them from the model (taking the set of all possible worlds to be those R-related to the actual world) without affecting the truth values (at the actual world) of any sentence. This generated submodel will be a simple S5 model, with a universal accessibility relation. But as soon as one enriches the language with other modal and epistemic operators, the situation changes. In the theory with two or more agents, even if one assumes that all agents have only true beliefs, the full S5 logic will not be preserved. The idealizing assumption will imply that Alice's beliefs coincide with her knowledge (in the actual world), and that Bob's do as well, but it will not follow that Bob knows (in the actual world) that Alice's beliefs coincide with her knowledge. To validate the full S5 logic, in the multiple agent theory, we need to assume that it is not just true, but common knowledge that everyone has only true beliefs. This stronger idealization is needed to reconcile the partition models, used in both game theory and in distributed systems theory, with the general theory that allows for a distinction between knowledge and belief. But even in a context in which one makes the strong assumption that it is common knowledge that no one is in error about anything, the possible divergence of knowledge and belief, and the failure of the S5 principles to be *necessarily* true will show itself when the language of knowledge and common knowledge is enriched with non-epistemic modal operators, or in semantic models that represent the interaction of epistemic and non-epistemic concepts. In game theory, for example, an adequate model of the playing of a game must represent not just the epistemic possibilities for each of the players, but also the capacities of players to make each of the choices that are open to that player, even when it is known that the player will not make some of those choices. One might assume that it is common knowledge that Alice will act rationally in a certain game, and it might be that it is known that Alice would be acting irrationally if she chose option X. Nevertheless, it would distort the representation of the game to deny that Alice has the option of choosing action

actual world of the model. In such theories, truth in all worlds in all models may diverge from truth in all models, allowing for semantic models of logics that fail to validate the rule of necessitation.

X, and the counterfactual possibility in which she exercises that option may play a role in the deliberations of both Alice and the other players, whose knowledge that Alice will not choose option X is based on their knowledge of what she knows would happen if she did. So even if one makes the idealizing assumption that all agents have only true beliefs, or that it is common belief that everyone's beliefs are true, one should recognize the more general structure that distinguishes belief from knowledge, and that distinguishes both of these concepts from subjective indistinguishability. In the more general structure that recognizes these distinctions, the epistemic accessibility relation is underdetermined by the other relations.

Minimal and Maximal Extensions

So our task is to say more about how to extend the relation D of doxastic accessibility to a relation R of epistemic accessibility. We know, from the assumption that knowledge implies belief, that in any model meeting our basic conditions on the relation between knowledge and belief, R will be an extension of D (for all x and y, if xDy, then xRy), and we know from the assumption that knowledge implies truth that the extension will be to a reflexive relation. We know by the assumption that belief is strong belief (belief that one knows) that R coincides with D, within the D-set (for all x and y, if xDx, then xRy if and only if xDy). What remains to be said is what determines, for a possible world x that is *outside* of a D-set, which other possible worlds outside that D-set are epistemically accessible to x. If some of my beliefs about what I know are false, what can be said about other propositions that I think that I know?

The assumptions of the neutral theory put clear upper and lower bounds on the answer to this question, and two ways to specify R in terms of the other resources of the model are to make the minimal or maximal extensions. The *minimal* extension of D would be the reflexive closure of D. On this account, the set of epistemically possible worlds for a knower in world x will be the set of doxastically accessible worlds, plus x. To make this minimal extension is to adopt the true belief analysis of knowledge, or in case one is making the internalist assumptions about justified belief, it would be to adopt the justified true belief analysis. The logic of true belief, S4.4, is stronger than S4.2, but weaker than S5.[16] The true belief analysis has its defenders, but most will want to impose stronger conditions on knowledge, which in our setting means that we need to go beyond the minimal extension of R.

It follows from the positive and negative introspection conditions for belief that for any possible world x, all worlds epistemically accessible to x will be subjectively indistinguishable from x (for all x and y, if xRy, then xEy) and this sets the upper bound on the extension of D to R. To identify R with the *maximal* admissible

[16]See the appendix for a summary of all the logics of knowledge discussed, their semantics, and the relationships between them.

extension is to define it as follows: xRy $=_{df}$ either (xDx and xDy) or (not xDx and xEy). This account of knowledge allows one to know things that go beyond one's internal states only when *all* of one's beliefs are correct. The logic of this concept of knowledge, S4F, is stronger than S4.2, but weaker than the logic of the minimal extension, S4.4. The maximal extension would not provide a plausible account of knowledge in general, but it might be the appropriate idealization for a certain limited context. Suppose one's information all comes from a single source (an oracle), who you presume, justifiably, to be reliable. Assuming that all of its pronouncements are true, they give you knowledge, but in possible worlds in which any one of its pronouncements is false, it is an unreliable oracle, and so nothing it says should be trusted. This logic, S4F, has been used as the underlying logic of knowledge in some theoretical accounts of a nonmonotonic logic. Those accounts don't provide an intuitive motivation for using this logic, but I think a dynamic model, with changes in knowledge induced by a single oracle who is presumed to be reliable, can provide a framework that makes intuitive sense of these nonmonotonic theories.[17]

Belief Revision and the Defeasibility Analysis

Any attempt to give an account of the accessibility relation for knowledge that falls between the minimal and maximal admissible extensions of the accessibility relation for belief will have to enrich the resources of the theory. One way to do this, a way that fits with one of the familiar strategies for responding to the Gettier counterexamples to the justified true belief analysis, is to add to the semantics for belief a theory of belief revision, and then to define knowledge as belief (or justified belief) that is stable under any potential revision by a piece of information that is in fact true. This is the defeasibility strategy followed by many of those who responded to Gettier's challenge: the idea was that the fourth condition (to be added to justified true belief) should be a requirement that there be no "defeater"—no true proposition that, if the knower learned that it was true, would lead her to give up the belief, or to be no longer justified in holding it.[18] There was much discussion in the post-Gettier literature, about exactly how defeasibility should be defined, but in the context of our idealized semantic models, supplemented by a semantic version of the standard belief revision theory, a formulation of a defeasibility analysis of knowledge is straightforward. First, let me sketch the outlines of the so-called AGM theory of belief revision,[19] and then give the defeasibility analysis.

[17] See Schwarz and Truszczyński (1992).

[18] See Lehrer and Paxson (1969) and Swain (1974) for two examples.

[19] See Gärdenfors (1988) for a survey of the basic ideas of the AGM belief revision theory, and Grove (1988) for a semantic formulation of the theory.

The belief revision project is to define, for each belief state (the prior belief state), a function taking a proposition (the potential new evidence) to a posterior belief state (the state that would be induced in one in the prior state by receiving that information as one's total new evidence). If belief states are represented by sets of possible worlds (the doxastically accessible worlds), and if propositions are also represented by sets of possible worlds, then the function will map one set of worlds (the prior belief set) to another (the posterior belief set), as a function of a proposition. Let **B** be the set representing the prior belief state, φ the potential new information, and **B**(φ) the set representing the posterior state. Let **E** be a superset of **B** that represents the set of all possible worlds that are potential candidates to be compatible with some posterior belief state. The formal constraints on this function are then as follows: (1) **B** (φ) → φ (the new information is believed in the posterior belief state induced by that information). (2) If φ ∩ **B** is nonempty, then **B** (φ) = φ ∩ **B** (If the new information is compatible with the prior beliefs, then nothing is given up—the new information is simply added to the prior beliefs.). (3) **B**(φ) is nonempty if and only if φ ∩ **E** is non-empty (the new information induces a consistent belief state whenever that information is compatible with the knower being in the prior belief state. and only then). (4) If **B** (φ) ∩ ψ is nonempty, then **B** (φ ∩ ψ) = **B** (φ) ∩ ψ. The fourth conditions is the only one that is not straightforward. What it says is that if ψ is compatible, not with Alice's *prior* beliefs, but with the *posterior* beliefs that she would have if she learned φ, then what Alice should believe upon learning the *conjunction* of φ and ψ should be the same as what she would believe if she first learned φ, and then learned ψ. This condition can be seen as a generalization of condition (2), which is a modest principle of methodological conservativism (Don't give up any beliefs if your new information is compatible with everything you believe). It is also a kind of path independence principle. The order in which Alice receives two compatible pieces of information should not matter to the ultimate belief state.[20]

To incorporate the standard belief revision theory into our models, add, for each possible world x, and for each agent i, a function that, for each proposition φ, takes i's belief state in x, $B_{x,i} = \{y: xD_iy\}$, to a potential posterior belief state, $B_{x,i}(\varphi)$. Assume that each of these functions meets the stated conditions, where the set **E**, for the function $B_{x,i}$ is the set of possible worlds that are subjectively indistinguishable from x to agent i. We will also assume that if x and y are subjectively indistinguishable to i, then i's belief revision function will be the same in x as it is in y. This is to extend the positive and negative introspection assumptions

[20]The third principle is the least secure of the principles; there are counterexamples that suggest that it should be given up. See Stalnaker (1994) for a discussion of one. The defeasibility analysis of knowledge can be given with either the full AGM belief revision theory, or with the more neutral one that gives up the fourth condition.

to the agent's belief revision policies. Just as she knows what she believes, so she knows how she would revise her beliefs in response to unexpected information.[21]

We have added some structure to the models, but not yet used it to interpret anything in the object language that our models are interpreting. Suppose our language has just belief operators (and not knowledge operators) for our agents, and only a doxastic accessibility relation, together with the belief revision structure, in the semantics The defeasibility analysis suggests that we might add, for knower i, a knowledge operator with the following semantic rule: $K_i\varphi$ is true in world x iff $B_i\ \varphi$ is true in x, and for any proposition ψ that is true in x, $\mathbf{B}_{x,i}(\psi) \rightarrow \varphi$. Alice knows that φ if and only if, for any ψ that is true, she would still believe that φ after learning that ψ. Equivalently, we might define an epistemic accessibility relation in terms of the belief revision structure, and use it to interpret the knowledge operator in the standard way. Let us say that $xR_i\ y$ if and only if there exists a proposition φ such that $\{x, y\} \subseteq \varphi$, and $y \notin \mathbf{B}_{x,i}(\varphi)$. The constraints imposed on the function $\mathbf{B}_{x,i}$ imply that this relation will extend the doxastic accessibility relation D_i, and that it will fall between our minimal and maximal constraints on this extension. The relation will be transitive, reflexive, and strongly convergent, and so meet all the conditions of our basic theory. It will also meet an additional condition: it will be weakly connected (if xRy and xRy, then either yRz, or zRy). This defeasibility semantics will validate a logic of knowledge, S4.3, that is stronger than S4.2, but weaker than either S4F or S4.4.[22]

So a nice, well behaved version of our standard semantics for knowledge falls out of the defeasibility analysis, yielding a determinate account, in terms of the belief revision structure, of the way that epistemic accessibility extends doxastic accessibility. But I doubt that this is a plausible account of knowledge in general, even in our idealized setting. The analysis is not so demanding as the S4F theory, but like that theory, it threatens to let any false belief defeat too much of our knowledge, even knowledge of facts that seem unrelated. Consider the following kind of example: Alice take herself to know that the butler didn't do it, since she saw him in the drawing room, miles away from the scene of the crime, at the time

[21]It should be noted that even with the addition of the belief revision structure to the epistemic models I have been discussing, they remain static models. A model of this kind represents only the agents' beliefs at a fixed time, together with the policies or dispositions to revise her beliefs that she has at that time. The model does not represent any actual revisions that are made when new information is actually received. The models can be enriched by adding a temporal dimension to represent the dynamics, but doing so requires that the knowledge and belief operators be time indexed, and that one be careful not to confuse belief changes that are changes of mind with belief changes that result from a change in the facts. (I may stop believing that the cat is on the mat because I learn that what I thought was the cat was the dog, or I may stop believing it because the cat gets up and leaves, and the differences between the two kinds of belief change are important).

[22]In game theoretic models, the strength of the assumption that there is common knowledge of rationality depends on what account one gives of knowledge (as well as on how one explains rationality). Some backward induction arguments, purporting to show that common knowledge of rationality suffices to determine a particular course of play (in the centipede game, or the iterated prisoners' dilemma, for example) can be shown to work with a defeasibility account of knowledge, even if they fail on a more neutral account. See Stalnaker (1996).

of the murder (or so she thinks). She also takes herself to know there is zucchini planted in the garden, since the gardener always plants zucchini, and she saw the characteristic zucchini blossoms on the vines in the garden (or so she thinks). As it happens, the gardener, quite uncharacteristically, failed to plant the zucchini this year, and coincidentally, a rare weed with blossoms that resemble zucchini blossoms have sprung up in its place. But it really was the butler that Alice saw in the drawing room, just as she thought. Does the fact that her justified belief about the zucchini is false take away her knowledge about the butler? It is a fact that either it wasn't really the butler in the drawing room, or the gardener failed to plant zucchini. Were Alice to learn just this disjunctive fact, she would have no basis for deciding which of her two independent knowledge claims was the one that was wrong. So it seems that, on the simple defeasibility account, the disjunctive fact is a defeater. The fact that she is wrong about one of her knowledge claims seems to infect other, seemingly unrelated claims. Now it may be right that if Alice was in fact reliably informed that one of her two knowledge claims was false, without being given any information about which, she would *then* no longer know that it was the butler that she saw. But if the mere fact that the disjunction is true were enough to rob her of her knowledge about the butler, then it would seem that almost all of Alice's knowledge claims will be threatened. The defeasibility account is closer than one might have thought to the maximally demanding S4F analysis, according to which we know nothing except how things seem to us unless we are right about everything we believe.

I think that one might plausibly defend the claim that the defeasibility analysis provides a *sufficient* condition for knowledge (in our idealized setting), and so the belief revision structure might further constrain the ways in which the doxastic accessibility relation can be extended to an epistemic accessibility relation. But it does not seem to be a plausible *necessary* and sufficient condition for knowledge. In a concluding section, I will speculate about some other features of the relation between a knower and the world that may be relevant to determining which of his true beliefs count as knowledge.

The Causal Dimension

What seems to be driving the kind of counterexample to the defeasibility analysis that I have considered is the fact that, on this analysis, a belief with a normal and unproblematic causal source could be defeated by the fact that some different source had delivered misinformation about some independent and irrelevant matter. Conditions were normal with respect to the explanation of Alice's beliefs about the butler's presence in the drawing room. There were no anomalous circumstances, either in her perceptual system, or in the conditions in the environment, to interfere with the normal formation of that belief. This was not the case with respect to the explanation of her belief about what was planted in the garden, but that does not seem, intuitively, to be relevant to whether her belief about the butler constituted knowledge. Perhaps the explanation of epistemic accessibility, in the case where conditions are not fully normal, and not all of the agent's beliefs are true, should focus more on the causal sources of beliefs, rather than on how agents would

respond to information that they do not in fact receive. This, of course, is a strategy that played a central role in many of the responses to the Gettier challenge. I will describe a very simple model of this kind, and then mention some of the problems that arise in making the simple model even slightly more realistic.

Recall that we can formulate the basic theory of belief this way: a relation of subjective indistinguishability, for each agent, partitions the space of possibilities, and there will be a nonempty subset of each partition cell which is the set of worlds compatible with what the agent believes in the worlds in that cell. We labeled those worlds the normal one, since they are the worlds in which everything determining the agent's beliefs is functioning normally, all of the beliefs are true in those worlds, and belief and knowledge coincide. The problem was to say what the agent knows in the worlds that lie outside of the normal set. One idea is to give a more detailed account of the normal conditions in terms of the way the agent interacts with the world he knows about; we start with a crude and simple model of how this might be done. Suppose our agent receives his information from a fixed set of independent sources—different informants who send messages on which the agent's knowledge is based. The "informants" might be any kind of input channel. The agent might or might not be in a position to identify or distinguish different informants. But we assume that the informants are, in fact, independent in the sense that there may be a fault or corruption that leads one informant to send misinformation (or more generally, to be malfunctioning) while others are functioning normally. So we might index normal conditions to the informant, as well as to the agent. For example, if there are two informants, there will be a set of worlds that is normal with respect to the input channel for informant one, and an overlapping set that is normal for informant two. Possible worlds in which conditions are fully normal will be those in which all the input channels are functioning normally—the worlds in the intersection of the two sets.[23] This intersection will be the set compatible with the agent's beliefs, the set where belief and knowledge coincide. If conditions are abnormal with respect to informant one (if that information channel is corrupted) then while that informant may influence the agent's beliefs, it won't provide any knowledge. But if the other channel is uncorrupted, the beliefs that have it as their sole source will be knowledge. The formal model suggested by this picture is a simple and straightforward generalization of the S4F model, the maximal admissible extension of the doxastic accessibility relation. Here is a definition of the epistemic accessibility relation for the S4F semantics, where $\mathbf{E}(x)$ is the set of worlds subjectively indistinguishable from x (to the agent in question) and $\mathbf{N}(x)$ is the subset of that set where conditions are normal (the worlds compatible with what the agent believes in world x): xRy if and only if $x \in \mathbf{N}(x)$ and $y \in \mathbf{N}(x)$, or $x \notin \mathbf{N}(x)$ and $y \in \mathbf{E}(x)$. In the generalization, there is a finite set of normal-conditions properties, $\mathbf{N}^j$, one for each informant j, that each determines a subset of $\mathbf{E}(x)$, $\mathbf{N}^j(x)$, where conditions are functioning normally in the relation between that informant and the agent. The definition of R will say that the analogue of the

[23] It will be required that the intersection of all the normal-conditions sets be nonempty.

S4F condition holds for each $\mathbf{N}^j$. The resulting logic (assuming that the number of independent information channels or informants is unspecified) will be the same as the basic theory: S4.2.

Everything goes smoothly if we assume that information comes from discrete sources, even if the agent does not identify or distinguish the sources. Even when the agent makes inferences from beliefs derived from multiple sources, some of which may be corrupt and other not, the model will determine which of his true beliefs count as knowledge, and which do not. But in even a slightly more realistic model, the causal explanations for our beliefs will be more complex, with different sources not wholly independent, and deviations from normal conditions hard to isolate. Beliefs may have multiple interacting sources—there will be cases of overdetermination and preemption There will be problems about how to treat cases where a defect in the system results, not in the reception of misinformation, but from the failure to receive a message. (It might be that had the system been functioning normally, I would have received information that would have led me to give up a true belief.) And along with complicating the causal story, one might combine this kind of model with a belief revision structure, allowing one to explore the relation between beliefs about causal structure and policies for belief revision, and to clarify the relation between the defeasibility analysis and an account based on the causal strategy. The abstract problems that arise when one tries to capture a more complex structure will reflect, and perhaps help to clarify, some of the patterns in the counterexamples that arose in the post-Gettier literature. Our simple model avoids most of these problems, but it is a start that may help to provide a context for addressing them.

Appendix

To give a very concise summary of all the logics of knowledge I have discussed, and their corresponding semantics, I will list, first the alternative constraints on the accessibility relation, and then the alternative axioms. Then I will distinguish the different logics, and the semantic conditions that are appropriate to them in terms of the items on the lists.
Conditions on R:

(Ref)	$(x)xRx$
(Tr)	$(x)(y)(z)((xRy \ \& \ yRz) \rightarrow xRz)$
(Cv)	$(x)(y)(z)((xRy \ \& \ yRz) \rightarrow (\exists w)(yRw \ \& \ zRw))$
(SCv)	$(x)(\exists z)(y)(xRy \rightarrow yRz)$
(WCt)	$(x)(y)(z)((xRy \ \& \ xRz) \rightarrow (yRz \vee zRy))$
(F)	$(x)(y)\big(xRy \rightarrow ((z)(xRz \rightarrow yRz) \vee (z)(xRz \rightarrow zRy))$
(TB)	$(x)(y)(xRy \ \& \ x \neq y)(z)(xRz \rightarrow zRy)\big)$
(E)	$(x)(y)(z)((xRy \ \& \ xRz) \rightarrow yRz)$

Axioms:

(T)	$K\varphi \rightarrow \varphi$
(4)	$K\varphi \rightarrow KK\varphi$
(4.2)	$MK\varphi \rightarrow KM\varphi$
(4.3)	$(K(\varphi \rightarrow M\psi) \vee K(\psi \rightarrow M\varphi))$
(f)	$((M\varphi \& MK\psi) \rightarrow K(M\varphi \vee \psi))$
(4.4)	$((\varphi \& MK\psi) \rightarrow K(\varphi \vee \psi))$
(5)	$M\varphi \rightarrow KM\varphi$

The logics for knowledge we have considered, and semantic constraints on R relative to which they are sound and complete, are as follows: The logics are of increasing order of strength, the theorems of each including those of the previous logics on the list.

S4	K + T + 4	Ref + Tr		
S4.2	S4 + 4.2	Ref + Tr + SCv	OR	Ref + Tr + Cv
S4.3	S4 + 4.3	Ref + Tr + SCv + WCt	OR	Ref + Tr + WCt
S4F	S4 + f	Ref + Tr + F		
S4.4	S4 + 4.4	Ref + Tr + TB		
S5	S4 + 5	Ref + Tr + E		

In each of the logics of knowledge we have considered, from S4.2 to S4.4, the derived logic of belief, with belief defined by the complex operator MK, will be KD45. (In S4, belief is not definable, since in that logic, the complex operator MK does not satisfy the K axiom, and so is not a normal modal operator. In S5, belief and knowledge coincide, so the logic of belief is S5.) KD45 is K + D + 4 + 5, where D is $(K\varphi \rightarrow M\varphi)$. The semantic constraints are Tr + E + the requirement that the accessibility relation be *serial*: $(x)(\exists y)xRy$.

In a semantic model with multiple knowers, we can add a common knowledge operator, defined in terms of the transitive closure of the epistemic accessibility relations for the different knowers. For any of the logics, from S4 to S4.4, with the corresponding semantic conditions, the logic of *common* knowledge will be S4, and the accessibility relation will be transitive and reflexive, but will not necessarily have any of the stronger properties. If the logic of knowledge is S5, then the logic of common knowledge will also be S5, and the accessibility relation will be an equivalence relation.

References

Battigalli, P., & Bonanno, G. (1999). Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics, 53*, 149–225.

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. (1995). *Reasoning about knowledge*. Cambridge, MA: MIT Press.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge, MA: MIT Press.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis, 6*, 121–123.

Grove, A. (1988). Two modeling for theory change. *Journal of Philosophical Logic, 17*, 157–170.

Hintikka, J. (1962). *Knowledge and belief*. Ithaca: Cornell University Press.

Lehrer, K., & Paxson, T. (1969). Knowledge: Undefeated justified true belief. *The Journal of Philosophy, 66*, 225–237.

Lenzen, W. (1978). *Recent work in epistemic logic* (Acta philosophica Fennica, Vol. 30). Amsterdam: North-Holland.

Schwarz, G., & Truszczyński, M. (1992). Modal logic S4F and the minimal knowledge paradigm. In *Proceedings of the fourth conference on theoretical aspects of reasoning about knowledge* (pp. 184–198). Sam Mateo: Morgan Kaufmann Publishers, Inc.

Stalnaker, R. (1991). The problem of logical omniscience, I. *Synthese, 89*, 425–440. (Reprinted in Stalnaker (1999a), 240–254).

Stalnaker, R. (1994). What is a non-monotonic consequence relation? *Fundamenta Informaticae, 21*, 7–21.

Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy, 12*, 133–162.

Stalnaker, R. (1999a). *Context and content: Essays on intentionality in speech and thought*. Oxford: Oxford University Press.

Stalnaker, R. (1999b). "The Problem of Logical Omniscience II" in Stalnaker (1999a), 255–273.

Swain, M. (1974). Epistemic defeasibility. *The American Philosophical Quarterly, 11*, 15–25.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

Chapter 31
Sentences, Belief and Logical Omniscience, or What Does Deduction Tell Us?

Rohit Parikh

Introduction

In deductive reasoning, if ϕ is deduced from some set Γ, then ϕ is already implicit in Γ. But then how do we learn anything from deduction? That we do not learn anything is the (unsatisfying) answer suggested by Socrates in Plato's *Meno*. This problem is echoed in the problem of logical omniscience prominent in epistemic logic according to which an agent knows all the consequences of his/her knowledge. An absurd consequence is that someone who knows the axioms of Peano Arithmetic knows all its theorems.

Since knowledge presumes belief, the lack of closure of (actual) beliefs under deduction remains an important issue.

The post-Gettier (1963) literature has concentrated on the gap between justified true belief and knowledge, but has not concerned itself with what *belief* is. This question, or at least an analysis of sentences of the form *Jack believes that frogs have ears*, **has** been prominent in the philosophy of language. But even there, less attention has been paid to *how* we know what someone believes. This will turn out to be an important factor.

The question "*How we know what someone believes*" has, however, been addressed by Ramsey and de Finetti as well as by Savage in the context of decision theory and the foundations of subjective probability. Beliefs are revealed by the choices we make, the bets we accept and the bets we refuse. And among these choices are the choices of what to say and what to assent to. Of course the second kind of choice can only be made by humans, or at least by creatures possessing

R. Parikh (✉)
City University of New York, 365 Fifth Avenue, New York, NY 10016, USA
e-mail: rparikh@gc.cuny.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_31

language. But the first kind of choice is perfectly open to animals and to pre-lingual children.[1]

We argue that this way of thinking about beliefs not only allows us to address the issue of logical omniscience and to offer formal models of "inconsistent" beliefs, it also allows us to say something useful about Frege's problem, whether it is sentences or something else that we believe, and whether *believes* can be a binary relation at all.[2]

We start with a dialogue between Socrates, Meno, and Meno's slave boy from Plato's *Meno*. In this dialogue, Socrates carries on a conversation with the boy and asks him the area (space) of a square whose side is two. The boy correctly answers that the area is four. And then Socrates raises the question of *doubling* the area to eight. He wants to know what the side should be. The boy makes two conjectures, first that the side of the larger square should also be double (i.e., four) – but that yields an area of sixteen, twice what is wanted. The boy's second guess is that the side should be three – but that yields an area of nine, still too large.

> **Socrates:** Do you see, Meno, what advances he has made in his power of recollection? He did not know at first, and he does not know now, what is the side of a figure of eight feet: but then he thought that he knew, and answered confidently as if he knew, and had no difficulty; now he has a difficulty, and neither knows nor fancies that he knows.
> **Meno:** True.
> **Socrates:** Is he not better off in knowing his ignorance?
> **Meno:** I think that he is.
> **Socrates:** If we have made him doubt, and given him the "torpedo's shock," have we done him any harm?
> **Meno:** I think not.
> **Socrates:** We have certainly, as would seem, assisted him in some degree to the discovery of the truth; and now he will wish to remedy his ignorance, but then he would have been ready to tell all the world again and again that the double space should have a double side.

Now Socrates suggests that the diagonal of the smaller square would work as the side of the larger square and the boy agrees to this. We continue with the quotation:

> **Socrates:** And that is the line which the learned call the diagonal. And if this is the proper name, then you, Meno's slave, are prepared to affirm that the double space is the square of the diagonal?
> **Boy:** Certainly, Socrates.

[1] In Ruth Marcus (1990, 1995) describes a man and his dog in a desert, deprived of water and thirsty. When they both react the same way to a mirage, it is hard to deny that they both have the same false belief that they are seeing water. The fact that animals can experience mirages would seem to be substantiated by the fact that the Sanskrit word for a mirage is *mrigajal* which literally means 'deer water,' faux water which deer pursue to their deaths. Frans de Waal makes a much more detailed case for animal intentionality in Waal (2005).

[2] In Levi (1997), Isaac Levi considers the doxastic commitment we have to *try to achieve* logical closure of our beliefs, even when, as he admits, we cannot actually achieve such logical closure. I am sympathetic to Levi's requirement, but in this paper, my concern is to develop a theory of actual beliefs rather than of doxastic commitments. The latter are less problematic from a purely logical point of view. If the agent's full beliefs are consistent, then his doxastic commitments will satisfy a suitable modal logic, perhaps the logic KD4.

Socrates: What do you say of him, Meno? Were not all these answers given out of his own head?
Meno: Yes, they were all his own.
Socrates: And yet, as we were just now saying, he did not know?
Meno: True.
Socrates: But still he had in him those notions of his – had he not?
Meno: Yes.
Socrates: Then he who does not know may still have true notions of that which he does not know?
Meno: He has.

—

Here Socrates appears to be arguing that the beliefs which Meno's slave could be brought to have via a discussion were beliefs which he *already* had.[3] There is of course a tension in this line of argument, for if this is so, then deduction would appear to be dispensable, and yet, leading the boy to make deductions was Socrates' own method in bringing him to a new state of mind.

A conclusion similar to that of Socrates also follows from the Kripke Semantics which has become a popular tool for formalizing logics of knowledge. The semantics for the logic of knowledge uses Kripke structures with an accessibility relation R, typically assumed to be reflexive, symmetric, and transitive. If we are talking about belief rather than knowledge, then R would be serial, transitive, and Euclidean. Then some formula ϕ is said to be believed (or known) at state s iff ϕ is true at all states R-accessible from s. Formally,

$$s \models B(\phi) \text{ iff } (\forall t)(sRt \rightarrow t \models \phi)$$

If follows immediately that if a formula is logically valid then it is true at all states and hence it is both known and believed. Moreover, if ϕ and $\phi \rightarrow \psi$ are believed at s then both are true at all t such that sRt, and hence ψ is true at all such t. Thus ψ is also believed at s. Moreover, a logically inconsistent formula can be neither known nor believed. Aumann's (1976) semantics uses partitions rather than Kripke structures, but is known to be equivalent (Fagin et al. 1995) and suffers from the same difficulty.

There have been suggestions, (Fagin et al. 1995; Moses 1988) that these 'problems' can be dealt with by using tools like *awareness*, *impossible possible worlds*, or by referring to *computational complexity*. In our view, these methods do not address the fundamental problem, namely *what is it that we are trying to model?* In order to have a theory of knowledge, we need to have some criterion for "measuring" what someone knows. Just to start with a logic which yields logical

[3]This impression is surely strengthened by the remarks which Socrates makes elsewhere in *Meno* to the effect that if knowledge was always present then the soul must be eternal. *The soul, then, as being immortal, and having been born again many times, and having seen all things that exist, whether in this world or in the world below, has knowledge of them all; and it is no wonder that she should be able to call to remembrance all that she ever knew about virtue, and about everything.*

omniscience, and then fiddling with it by various methods, leaves aside the question of what our goal is. Unless we have a clear sight of the goal we are not likely to reach it.

For instance, we do not know if the Goldbach conjecture is true, but if it is true, then it is necessarily true. Surely we cannot argue that we do not know which only because we are not *aware* of it. That is surely not the problem. Nor can computational complexity help us, because *if* it is true, then there is a sound formal system which has it as an axiom, and then a proof of the Goldbach conjecture is going to be very fast in that system. If it is false, then of course the falsity is provable in arithmetic as it only requires us to take notice of a particular even number which is not the sum of two (smaller) primes.

The issue of computational complexity can only make sense for an infinite family of questions, whose answers may be undecidable or at least not in polytime. But for *individual* mathematical questions whose answers we do not know, the appeal to computational complexity misses the issue.

In sum, our goal is to first answer the question, *How do we know what people believe?* and then try to develop a theory of what people do believe.

Inconsistent Beliefs

Moreover, beliefs are not always consistent, so one could say, *thank heaven for a lack of logical omniscience!* A person, some of whose beliefs conflict with others, and who believes all logical consequences of her belief, would end up believing *everything!*

The following two examples are from Daniel Kahneman's Nobel lecture, (2002)

> *A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" Almost everyone reports an initial tendency to answer "10 cents" because the sum $1.10 separates naturally into $1 and 10 cents, and 10 cents is about the right magnitude. Frederick found that many intelligent people yield to this immediate impulse: 50 % (47/93) of Princeton students, and 56 % (164/293) of students at the University of Michigan gave the wrong answer. Clearly, these respondents offered a response without checking it.*
>
> ———
>
> *Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice and also participated in antinuclear demonstrations.*
> *#6* Linda is a bank teller
> *#8* Linda is a bank teller and active in the feminist movement
> *89 % of respondents rated item #8 higher in probability than item #6.*

But the set of bank tellers who are active in the feminist movement is a proper subset (perhaps even a rather small subset) of the set of all bank tellers, so #8 *cannot* have higher probability[4] than #6.

[4]The "conjunction fallacy" is committed when someone assigns higher probability to a conjunction than to one of the conjuncts. Gigerenzer (1996), Levi (2004) and Hintikka (2004) all dispute that the 89 % of people who responded the way indicated were actually committing the conjunction

Clearly human states of belief are not usually consistent. Nor are they usually closed under logical inference. Various researchers including ourselves have looked into this issue (Fagin et al. 1995; Gaifman 2004; Parikh 1987, 1995; Stalnaker 1999). There is also more recent work by Artemov and Nogina (2005), and by Fitting (2004), aimed towards the logic of proofs.

And yet we do operate in the world without getting into trouble, and when we discover an inconsistency or incompleteness in our thinking, we remove it, either by adding some beliefs which we did not have before, or by deleting some which we had. What we need is a formal account of this activity and a more general representation of beliefs than is afforded by the current Kripke semantics. In particular, logically omniscient belief states need to be represented as a *proper subset* of all belief states.

The proposal we make here draws on the work of Ramsey (1931), de Finetti (1937), Hayek (1948), and Savage (1954) with echoes from Marcus (1990), Marcus (1995), Millikan (2006), and Whyte (1990).[5] According to this view, an agent's beliefs are revealed by the *choices* which an agent makes, and while 'incoherent' choices are unwise, they are not inconsistent, for if they were, they would be *impossible*. For an agent to make such unwise, incoherent, choices is perfectly possible and is done all the time.

Imagine that Carol assigns probabilities of .3, .3, and .8 respectively to events $X, Y, X \cup Y$. One could say that these probabilities are *inconsistent*. But in fact nothing prevents Carol from *accepting* bets based on these probabilities. What makes them incoherent is that we can make Dutch book against Carol – i.e., place bets in such a way that no matter what happens, she will end up losing money.[6] Thus incoherent beliefs, on this account, are *possible*, (and hence consistent) but unwise. It will be important to keep in mind this distinction between inconsistency and incoherence.

We now look at some previous work which contains strong indications of the direction in which to proceed.

Hayek (1948) considers an isolated person acting over a period according to a preconceived plan. The plan

> *may, of course, be based on wrong assumptions concerning external facts and on this account may have to be changed. But there will always be a conceivable set of external events which would make it possible to execute the plan as originally conceived.*

fallacy. However, I assume that the dispute is about the interpretation of this particular experiment, and that these three writers would not dispute the *more general point* that people do sometimes reason incorrectly.

[5] However, unlike Ramsey et al., we shall not try to explain probabilistic belief.

[6] For instance we can bet \$3 on X, \$3 on Y, and \$2 against $X \cup Y$. If either X or Y happens, we earn \$7 (at least), and lose (at most) \$5. If neither happens, we gain \$8 and lose \$6, so that we again make a profit – and Carol makes a loss.

Beliefs are related here to an agent's plans, and it is implicit in Hayek's view that the agent believes that the world is such that his plans will work out (or have a good chance of doing so).

But note that the belief states which are implicit in plans are more general than the belief states which correspond to Kripke structures, and the first may be incoherent and/or logically incomplete. An agent may plan to buy high and sell low, and expect to make money. It is not possible for the agent to actually do so, but it is perfectly possible for the agent to have such a plan.

Animal Beliefs

Now, for a creature to have a plan it is not necessary that the plan be formulated explicitly in language, or even that the planner *has* a language. It is perfectly possible for an animal to have a plan and to a smaller extent, it is also possible for a pre-lingual child to engage in deliberate behaviour which is plan-like.

This is an important point made explicitly by Marcus (1995), Searle (1994), and also fairly clear in Millikan (2006), that we ought not to limit states like belief and desire to language using creatures, i.e., to adult humans and older children. Frans de Waal is even more emphatic on this point.[7] We should also attribute some kinds of intentional states to higher animals and to pre-lingual children. Going back further, Hume (1988) is quite explicit on this point:

> *Next to the ridicule of denying an evident truth is that of taking much pains to defend it; and no truth appears to me more evident, than that beasts are endow'd with thought and reason as well as men.*

Ramsey (1990) has a related comment.[8]

> *It is for instance possible to say that a chicken believes a caterpillar of a certain sort to be poisonous,and mean by that merely that it abstains from eating such caterpillars on account of unpleasant experiences connected with them.*

And so does Millikan (2006).

> *Reasoning is just trial and error in thought. Dennett (1996) calls animals capable of trial and error in thought "Popperian." The reference is to Popper's remark that* it is better to let one's hypotheses die in one's stead. *The Popperian animal is capable of thinking*

[7] "It wasn't until an ape saved a member of our own species that there was public awakening to the possibility of nonhuman kindness. This happened on August 16, 1996 when an eight-year old female gorilla named Binti Jua helped a three-year-old boy who had fallen eighteen feet into the primate exhibit at Chicago's Brookfield Zoo. Reacting immediately, Binti scooped up the boy and carried him to safety." De Waal is quite disdainful of Katherine Hepburn's remark in *The African Queen*: "Nature, Mr. Allnut, is what we are put in this world to rise above."

[8] Of course we need not and should not attribute to the chicken the specific belief that such caterpillars are *poisonous.* Davidson (1982) is right on this particular point. But we *can* attribute to it the belief that eating them will lead to bad consequences.

hypothetically, of considering possibilities without yet fully believing or intending them. The Popperian animal discovers means by which to fulfill its purposes by trial and error with inner representations. It tries things out in its head, which is, of course, quicker and safer than trying them out in the world. It is quicker and safer than either operant conditioning or natural selection. One of many reasonable interpretations of what it is to be rational is that being rational is being a Popperian animal. The question whether any non-human animals are rational would then be the question whether any of them are Popperian.

Finally Whyte (1990) suggests that we can even define truth in this way. He appeals to Ramsey's principle (R):

(R) *A belief's truth condition is that which guarantees the fulfilment of any desire by the action which that belief and desire would combine to cause.*

Defining Belief

However, we need not address the issue of truth here. Brandom (1994) subjects Whyte to some criticism, but it only has to do with whether the criterion of truth is adequate. We are only interested here in a *representation* of belief. Such a representation must accommodate all the following groups: language possessing, logically omniscient humans[9]; language possessing but fallible humans; and non-lingual creatures like animals and very young children.[10]

However, if beliefs are not (always) expressed by sentences, or coincide with propositions expressed by sentences, then we need another way of representing belief states, and then relate language-oriented belief states to some specific species of such belief states.

We shall consider two kinds of beliefs. Non-linguistic beliefs which may also be possessed by animals, and linguistic beliefs which can only be possessed by

[9]Of course I do not personally know any logically omniscient humans, but in a limited context it is possible for a human to show full logical competence. Suppose that p stands for *Pandas live in Washington DC*, q stands for *Quine was born in Ohio*, and r stands for *Rabbits are called* gavagai *at Harvard.* Suppose that Jill believes that p is true and that q and r have the same truth values. Then she is allowing two truth valuations, $v = (t, t, t)$, and $v' = (t, f, f)$. Given a formula ϕ on p, q, r in disjunctive normal form, she can evaluate $v(\phi)$ and $v'(\phi)$. If both are t she can say that she believes ϕ. If both are f, she disbelieves ϕ, and otherwise she is suspending judgment. Then Jill *will* be logically omniscient in this domain. But note that she will actually have to *make the calculations* rather than just sit back and say, "Now *do* I believe ϕ?" In fact if it so happens that ϕ is a complex formula logically equivalent to p, then ϕ represents the same proposition as p, and is therefore believed by Jill. And yet, Jill will not agree to ϕ *because* it is the same 'proposition' as p, but rather that she will agree to the *formula* ϕ whose truth value she has calculated. See also, Dennett (1985) p. 11.

[10]It is plausible that when a vervet monkey utters a leopard call, then it is saying, '*Ware leopard!*', but surely nothing in the behaviour of such monkeys would justify us to think that they might utter *If there were no leopards here, then this would be a wonderful spot to picnic.*

humans; adults and older children. Of course the last two groups will also have non-linguistic beliefs which must be somehow correlated with their linguistic beliefs.

Let $\mathcal{B}$ be the space (so far unspecified) of belief states of some agent. Then the elements of $\mathcal{B}$ will be correlated with the choices which the agent makes. Roughly speaking, if I believe that it is raining, I will take my umbrella, and if I believe that it is not raining, then I won't. What I believe is revealed by what I do.

But clearly the choice of whether to take my umbrella or not is correlated with my belief only if I don't want to get wet. So my preferences enter *in addition* to my beliefs.

Conventional theories like that of Savage (1954) take the (actual and potential) choices of an agent, assume that these choices satisfy certain axioms, and simultaneously derive both the agent's beliefs (her subjective probability) and the agent's preferences (expressed by a utility function). But it is known that agents which are rational in Savage's sense, i.e., obey his axioms, are not very common and so we would like to retain the *rough framework* without the inappropriate precision of Savage's theory.

So we will just assume that the agent has *some* space $\mathcal{P}$ of preferences, and that the choices are governed by the beliefs as well as the preferences. We will use $\mathcal{S}$ for the set of choice situations, and C for the set of choices. Thus the set $\{U, \neg U\}$ could be a choice situation, or an element of $\mathcal{S}$, (with U standing for *take the umbrella*); and both U and $\neg U$ are elements of C. We could say very roughly that an agent who prefers not to get wet will choose U from $\{U, \neg U\}$ iff she believes that it is raining.

An agent who does have language can also be subjected to a purely linguistic choice. If asked *Do you think it is raining?* the agent may choose from the set *{Yes, No, Not sure}*. And it is going to be *usually* the case that the agent will choose U in the situation, $\{U, \neg U\}$ iff she chooses *Yes* in the situation where she hears *Do you think that it is raining?* But this is not a *logical* requirement, only a pragmatic one.

We will say that an agent *endorses* (agrees with) a sentence ϕ iff she chooses *Yes* when asked, *Do you think* ϕ*?*, and *denies* (or disagrees with) ϕ iff she chooses *No*. She may also choose *Not sure*, in which case of course she neither endorses nor denies.

Note that nothing prevents an agent from endorsing ϕ as well as $\neg\phi$, but few agents are that 'out of it'. But of course an agent may endorse ϕ, $\phi \rightarrow \psi$, and either deny ψ or at least fail to endorse ψ. We will say in the first case that the agent is logically incoherent, and in the second that the agent is an incomplete reasoner – or simply incomplete.

Given the fact that $(\phi \wedge \neg\phi) \rightarrow \psi$ for arbitrary ψ is a tautology, it is obvious that an agent who is logically incoherent, but also logically complete, and who endorses both ϕ and $\neg\phi$, will end up endorsing *everything*. Fortunately, most of us, though we *are* logically incoherent, tend also to be incomplete. If we endorse the statements that all men are equal, that Gandhi and Hitler are men, and that Gandhi and Hitler are not equal, we are still not likely to agree that pigs fly – which we would if we were also *complete.*

As we have made clear, elements of $\mathcal{B}$ cannot be identified with propositions, for an agent may agree to one sentence expressing a proposition and disagree (or

not agree with) another sentence expressing the same proposition. An agent in some state $b \in \mathcal{B}$ may agree with *Hesperus is bright this evening*, while disagreeing with *Phosphorus is bright this evening*. But surely we *may* think of an agent speaking some language $\mathcal{L}$ as agreeing with or endorsing a particular *sentence* ϕ.

Definition 1. A belief state b is *incomplete* when there are sentences $\phi_1, \ldots, \phi_n$ which are endorsed in b, ψ follows logically from $\phi_1, \ldots, \phi_n$, and ψ is not endorsed in b.

A belief state b is *incoherent* when there are sentences $\phi_1, \ldots, \phi_n$ which are endorsed in b, ψ follows logically from $\phi_1, \ldots, \phi_n$, and ψ is denied in b.

There is of course an entirely different sort of incoherence which arises when an agent's linguistic behaviour does not comport with his choices. Suppose an agent who prefers not to get wet, *says* that it is raining, but does not take her umbrella, then she may well be quite coherent in her linguistic behaviour, but her linguistic behaviour and her non-linguistic choices have failed to match. Whether we then conclude that the agent is irrational, is being deceptive, or does not know the language, will depend on the weight of the evidence.

We can easily see now how deduction changes things. Meno's slave was initially both incoherent and incomplete. He believed that a square of side four had an area of eight. Since he knew some arithmetic and some geometry, his belief state was not actually coherent. At the end of the conversation with Socrates, he came to endorse the sentence, *The square whose side is the diagonal of a square of side two, has an area of eight*. It is not quite clear what practical application this information had for him in this case, but surely carpenters carrying out deductions of the same kind will manage to make furniture which does not collapse and bears the load of people and objects.

Beliefs can change not only as a result of deductions, as those Meno's slave did, they can also change as a result of experience, e.g., raindrops falling on your head, or as the result of hearing something, like *It is raining.*

A Philosophical Aside

In this paper I shall avoid taking sides in philosophical disputes, but one thing does need to be mentioned. The representational account of belief seems simply wrong to me. We *can* certainly think of beliefs as being stored in the brain in some form and called forth as needed, but when we think of the details, we can soon see that the stored belief model is too meager to serve. I will offer an analogy.

Suppose Jennifer owns a bookstore. If someone calls her and asks if she has Shakespeare's *Hamlet*, she will look on her shelves, and answer yes if *Hamlet* is in her *store*. But suppose now that someone asks her if she has Shakespeare's *Tragedies*. These include *Hamlet* of course, but also *Macbeth*, *Romeo and Juliet* and *King Lear*. If they are stored alphabetically by title, then they will be in different locations and they won't be in her store as *one item*. But she can *create*

the set and ship it out as one item to the customer. Did she have the set when the customer called? Surely yes. It would sound silly to say to the customer, "I do have *Hamlet, Macbeth, Romeo and Juliet* and *King Lear*, but unfortunately I don't have Shakespeare's *Tragedies*".

Let us go one step further. Suppose she actually only has a CD containing the files for the four plays, a printer, and a binder. She can still create the set using what she has, even though at the time of the call she had no physical object corresponding to the noun phrase *Shakespeare's Tragedies*.

It is what she *has*, namely the CD, the printer, and the binder, and what she *can do*, namely print and bind, which together allow her to fulfill the order. There are elements of pure storage, and elements which are algorithmic which together produce the item in question. These two kinds of elements may not always be easy to separate. It is wiser just to concentrate on what she can supply to her customer.

It is the same, in my view, with beliefs. No doubt there are certain beliefs which are stored in some way, but there may be other equally valid beliefs which can be produced *on the spot* so to say, without having been there to start with. And note that if Jennifer has a large number of actual physical books, but also CDs for some, it may be easier for her to produce a book from a CD than to find a book which exists physically in the store.

If someone asks me if I believe that 232345456 is even, I can answer yes at once (since the number ends in a 6), even though that particular belief had never been stored. But if someone asks me the name of some classmate from high school, I might take a long time to remember. Retrieving from storage is *one* way to exhibit a belief, but not the only one, and often, not even the best one.

In this paper, I shall not assume any particular representation of beliefs, but deal with them purely in terms of how a person acts, and how the potential for acting in some way is affected by various kinds of update.

Some Technical Details

We assume given a space $\mathcal{B}$ for some agent whose beliefs we are considering. The elements of $\mathcal{B}$ are the belief states of that agent, and these are not assumed to be sentences in Mentalese although for some restricted purposes they could be. There are three important update operations on $\mathcal{B}$ coming about as a result of (i) events observed, (ii) sentences heard, and (iii) deductions made. Elements of $\mathcal{B}$ are also used to make *choices*. Thus in certain states of belief an agent may make the choice to take his umbrella and we could then say that the agent believes it is raining. Many human agents are also likely to make the choice to *say*, "I think it is raining and so I am taking my umbrella" but clearly only if the agent is English speaking. Thus two agents speaking different languages, both of whom are taking their umbrellas, but making different noises, have the same belief in one sense but a different one in another. Both of these will matter. Later on we shall look into the connection.

Deduction is an update which does not require an input from the outside. But it *can* result in a change in $\mathcal{B}$. Suppose for instance that Jane thinks it is clear, and is about to leave the building without an umbrella. She might say, "Wait, didn't I just see Jack coming in with a wet umbrella? It must be raining." The sight of Jack with a wet umbrella might not have caused her to believe that it was raining, perhaps she was busy with a phone call. But the memory of that wet umbrella may later cause a deduction to take place of the fact that it is raining.

Thus our three update operations are:

$$\mathcal{B} \times \mathcal{E} \to_e \mathcal{B}$$

A belief state gets revised by witnessing an event.

$$\mathcal{B} \times \mathcal{L} \to_s \mathcal{B}$$

A belief state gets revised through hearing a sentence.[11,12]

And hearing two logically equivalent sentences s, s' need not result in the same change occurring, although they may, in case the agent knows they are equivalent.[13]

$$\mathcal{B} \to_d \mathcal{B}$$

A deduction causes a change in the belief state (which we may sometimes represent as an *addition*).[14]

Here E is the set of events which an agent may witness and $\mathcal{L}$ is some language which an agent understands (i.e., uses successfully). The last map $\to_d$ for deduction is non-deterministic as an agent in the same state of belief may make one deduction or another.

Finally, we also have a space $\mathcal{S}$ of *choice sets* where an agent makes a particular choice among various alternatives. This gives us the map

$$\mathcal{B} \times \mathcal{S} \to_{ch} \mathcal{B} \times C$$

[11]Of course, hearing a sentence is also an event, but its effect on speakers of the language goes beyond just the event. It is this second part which falls under $\to_s$.

[12]Even a dog may revise its state of belief on hearing *Sit!*, see for instance Parikh and Ramanujam (2003). Note also that if the sentence heard is inconsistent with one's current beliefs *and* one notices the inconsistency, then some theory like that in Alchourron et al. (1985) may need to be deployed.

[13]Clearly Lois Lane will react differently to the sentences *Superman flew over the Empire State Building*, and *Clark Kent flew over the Empire State Building.* Similarly, Kripke's Pierre (1979) will react differently to the questions, *Would you like a free trip to Londra?* and *Would you like a free trip to London?* Indeed in the second case he might *offer to pay* in order not to go to London!

[14]See, however, Alchourron et al. (1985) where more complex kinds of reactions to sentences heard are described.

An agent in a certain belief state makes a choice among various alternatives, and may arrive at a different state of belief *after* making that choice.

If we want to explicitly include preferences, we could write,

$$\mathcal{B} \times \mathcal{P} \times \mathcal{S} \rightarrow_{ch} \mathcal{B} \times C$$

While $\mathcal{S}$ is the family of choice sets, C is the set of possible choices and $\mathcal{P}$ is *some* representation of the agent's preferences. Thus *{take umbrella, don't take umbrella}* is a choice set and an element of $\mathcal{S}$ but *take umbrella* is a choice and an element of C.

Example. Suppose that Vikram believes it is not raining. In that case he will be in a belief state b such that $ch(b, \{U, \neg U\}) = (b', \neg U)$. Given a choice between taking an umbrella or not, he chooses not to take the umbrella, and goes into state b'.

Suppose, however, that he looks out the window and sees drops of rain falling. Let r be the event of rain falling. Then the update operation $\rightarrow_e$ causes him to go into state c such that in state c he chooses U from $\{U, \neg U\}$. Thus $ch(c, \{U, \neg U\}) = (c', U)$

Note that state c will also have other properties beside choosing to take an umbrella. It may also cause him to say to others, "You know, it is raining," or to complain, "Gosh, does it always rain in this city?" There is no such thing as *the state* of believing that it is raining. Every belief state has many properties, most of them unrelated to rain.

Among the choices that agents make are choices to assent to or dissent from sentences. But there is no *logical* reason why an agent who assents to "It is raining" must take an umbrella or a raincoat. It is just the more pragmatic choice to take the umbrella when one says that it is raining, because otherwise either one gets wet, or one is suspected of insincerity. We could say that the agent believes the *sentence* "It is raining", and dis-believes the *proposition* that it is raining. But we would feel uncomfortable with such an account and might prefer to say that the agent is either lying or confused.

Of course an agent may prefer to get wet and in that case, saying "It is raining," and not taking an umbrella are perfectly compatible choices. This shows that an agent's preferences need to be taken into account when correlating the actions which an agent takes and what an agent believes. But we usually do not want to get wet and to make such choices, and usually we do not say what we do not believe. It does not work for us.

Thus our theory of an agent presupposes such a belief set $\mathcal{B}$, and appropriate functions $\rightarrow_e, \rightarrow_s, \rightarrow_d, \rightarrow_{ch}$. We can understand an agent (with some caveats) if what we *see* as the effects of these maps conforms to some theory of what an agent wants and what the agent thinks. And we succeed pretty well. *Contra* Wittgenstein (1958), we not only have a theory of what a lion wants, and what it means when it growls, we even have theories for bees and bats. Naturally these theories do not have the map $\rightarrow_s$ except with creatures like dogs or cats and some parrots (who not only "parrot" our words but understand them to some extent (Pepperberg 2004)).

Partial Similarity of Belief States

A familiar notion of similarity used in mathematics is that of *isomorphism*. If two structures $< X, R >$, $< Y, R' >$ are isomorphic, then it means that there is a 1–1 function f from X onto Y such that for all $a, b \in X$, $R(a, b)$ iff $R'(f(a), f(b))$. Isomorphic structures have the same logical properties.

The notion which will be relevant for us, however, will be the notion of *bi-simulation* (Johan van Benthem 1976; Milner 1989; Park 1981).

In the same context as above, we will say that $< X, R >$ and $< Y, R' >$ are *bi-similar* if there is a subset B (the bi-simulation) of $X \times Y$ such that X is the domain of B, Y is the co-domain of B, and we have the following conditions.

- If $B(a, b)$ and there is c such that $R(a, c)$, then there is a d such that $B(c, d)$ and $R'(b, d)$.
- If $B(a, b)$ and there is d such that $R'(b, d)$, then there is a c such that $B(c, d)$ and $R(a, c)$.

Intuitively we can say that if a, b are related by the bi-simulation, then for anything c which is related by R to a, there is something d which is related by R' to b, and vice versa. Moreover, c, d are also related by B.

Thus the structures $< X, R >$ and $< Y, R' >$ can imitate each other without having to be isomorphic. Isomorphism implies bi-similarity but not vice versa.

If we have richer structures, $< X, R, S >$ and $< Y, R', S' >$ then it may well happen that $< X, R >$ and $< Y, R' >$ are bi-similar, but $< X, R, S >$ and $< Y, R', S' >$ are not. Then the imitation will be partial.

It can thus happen that two belief structures $\mathcal{B}$ and $\mathcal{B}'$ can be bi-similar in *some* ways, without being bi-similar in every way. Two students in the same school may show similar behavior in the cafetaria, and be bi-similar in their cafetaria personae, but not be bi-similar in, say, their dating behavior. One could say then, somewhat loosely, that they have the same beliefs about food but not about girls. (I am assuming here that their preferences are the same.)

Similarly, Jack and Jill can have the same beliefs and preferences about ice cream, and exhibit similar behavior in ice cream choosing settings, but may have different beliefs about politics.

So far we have only talked about belief states as being indicated by choices. But of course many beliefs are expressed (or so we think) by sentences. We may suggest, with Stalnaker, that what we believe are propositions. Though I do not think such accounts can work, the fact that we find them appealing requires some explanation.

We shall now introduce two notions of belief, e-belief (which is choice-based) and i-belief (which is sentence-based and only appropriate for language users). These two notions will be used to sort out various belief puzzles.

The Setting

In our setting we imagine an observer o who is pondering on what some agent i believes. We assume (for convenience) that o thinks of a proposition expressed by a sentence as a set of possible worlds where that sentence is true, but that the observee i need not even have a language or a notion of truth. However, it is assumed that i does have some plans. Even if i is just a dog digging for a bone, o understands that i has a plan and roughly what that plan is. And we shall use this plan to make it possible for o to attribute beliefs to i.

We also assume that there is a context C which is the set of *relevant* possible worlds, and that worlds outside C, even though they are there, are not considered in deliberating about i's belief or beliefs. It is not assumed that i knows that there are worlds outside C; in some sense i lives inside C, but we will assume that o does know that there are possible worlds outside C. The purpose of the context is to avoid considering cases which are possible but strange, like the laws of physics failing, or Bush suddenly joining the Green party. A plan is assumed to work in *normal* circumstances, and an agent i is only required to have the sort of belief which would be enough for the plan to be carried out in normal circumstances.

For instance, chickens, when given a choice between drinking water and drinking mercury tend to prefer the latter, presumably because it looks more like water (is shinier) than water itself does. We certainly do not want to attribute to the chickens a belief that mercury is good to drink, merely that if you are thirsty, you should go for a shiny liquid.

So let P be i's plan at the moment, and let $\pi(P)$ be the set of worlds w in C such that the plan is *possible* at w.

There are two senses in which the plan may be possible. One is that the plan can actually be carried out. For instance for the dog digging for a bone, that possibility means that the ground is soft enough. The other sense is that the plan actually yields i the benefit which i wants; in this case, the actual presence of a bone.

So formally,

$$\pi(P) = \{w|\ w \in C \wedge w\ enables\ P\}.$$

Let ϕ be a sentence. Then $||\phi|| = \{w|w \models \phi\}$, the set of worlds where ϕ is true, is the *proposition* corresponding to the sentence ϕ. Note that if ϕ and ψ are logically equivalent, then $||\phi|| = ||\psi||$.

Definition 2. We will say that i *e-believes* ϕ, $B^i_e(\phi)$ if $\pi(P) \subseteq ||\phi||$. We will suppress the superscript i when it is clear from context.

It is obvious in terms of the semantics which we just gave that the statement "The dog e-believes that there is a bone where he is digging" will be true from *our* point of view. A Martian observing the same dog, but not having the notion of a bone (I assume there are no vertebrates on Mars) will obviously not assign the same belief to the dog; but the Martian may still have his own theory of the dog's belief state

which will allow him to predict the dog's behavior. In other words, the Marian will assign a space $\mathcal{B}'$ to the dog which will be bi-similar in relevant ways to the space $\mathcal{B}$ which *we* would assign to the dog.

It is easy to see that if an agent e-believes ϕ and ψ then the agent also e-believes $\phi \wedge \psi$ and that if the agent e-believes ϕ and $\phi \rightarrow \psi$ then the agent e-believes ψ.[15] Oddly enough, creatures which do not use language do not suffer from a lack of logical omniscience!

A lot of logic goes along with e-belief, but only within the context of a single plan. For instance, one may drive a colleague to the airport for a two week vacation in Europe and then forget, and arrange a meeting three days later at which this colleague's presence is essential. But within the context of a single (short) plan, consistency and logical omniscience will tend to hold. The situation is more complex with multiple plans. And there is nothing to prevent an agent from having one e-belief in one plan and another contradicting e-belief in another plan. It is pragmatic considerations – the logical counterpart of avoiding Dutch Book – which will encourage the agent i to be consistent and to use logical closure.

Suppose someone has a plan P consisting of, "If ϕ then do α, else do β" and another plan P' consisting of "If ϕ then do γ, else do δ". Now we find him doing α and also doing δ (we are assuming that the truth value of ϕ has not changed). We could accuse him of being illogical, but there is no need to appeal to logic. For he is doing Dutch book against himself.

Presumably he assumed[16] that $u(\alpha|\phi) > u(\beta|\phi)$ but $u(\alpha|\neg\phi) < u(\beta|\neg\phi)$. Thus given ϕ, α was better than β but with $\neg\phi$ it was the other way around. Similarly, $u(\gamma|\phi) > u(\delta|\phi)$, but $u(\gamma|\neg\phi) < u(\delta|\neg\phi)$. And that is why he had these plans. But then his choice of α, δ results in a loss of utility whether ϕ is true or not. If ϕ is true then he lost out doing δ and if ϕ is false, then he lost out doing α.

For a concrete example of this, suppose that on going out I advise you to take your umbrella, but fail to take mine. If it is raining, there will be a loss of utility for I will get wet. If it is not raining, there will be a loss of utility because you will be annoyed at having to carry an umbrella for no good reason. My choice that I advise you to take your umbrella, but fail to take mine, is not *logically* impossible. It just makes no pragmatic sense. A similar argument will apply if someone endorses ϕ, endorses $\phi \rightarrow \psi$ and denies ψ. If such a person makes plans comporting with these three conditions, then he will make choices which do not maximise his utility. Of course such arguments go back to Ramsey (1931) and Savage (1954).

[15] To see this, if $\pi(P) \subseteq ||\phi||$, and $\pi(P) \subseteq ||\psi||$ then clearly $\pi(P) \subseteq ||\phi|| \cap ||\psi|| = ||\phi \wedge \psi||$. The proof for the other case is similar using the fact that $||\phi \rightarrow \psi|| = (C - ||\phi||) \cup ||\psi||$. Since $\pi(P)$ is contained in $||\phi||$, it is disjoint from $C - ||\phi||$. Hence it can be contained in $(C - ||\phi||) \cup ||\psi||$ if and only if it contained in $||\psi||$.

[16] The use of the letter u for utility is not meant to suggest that we have a formal notion of utility in mind; only a rough one.

If we assume that over time, people learn to maximise their utility (they do not always but often do), then they will 'learn' a certain amount of logic and they will make certain obvious logical inferences.[17]

A little girl who was in kindergarten was in the habit of playing with her older sister in the garden every day when she came home. Once, her older sister was sick. So the little girl went to visit her sick sibling in her bedroom, and then, as usual, went out into the garden to play with her. Clearly the little girl had not yet learned to maximise her utility!

A Second Notion of Belief: Language Enters

We now define a second notion of belief which does *not* imply logical omniscience. This is a more self-conscious, language-dependent notion of belief.

For agents i who do have a language (assumed to be English from now on), their plan may contain linguistic elements. At any moment of time they have a finite stock of currently believed sentences. This stock may be revised as time passes. These agents may perform atomic actions from time to time, and also make observations which may result in a revision in their stock of believed sentences.[18]

Thus Lois seeing Superman in front of her will add the sentence "Superman is in front of me", to her stock, but, since she does not know that Clark Kent is Superman, she will *not* add the sentence "Clark Kent is in front of me". Someone else may add the sentence "I see the Evening Star", but not the sentence "I see the Morning Star" at 8 PM on a summer night. A person who knows that $ES = MS$, may add the sentence, "Venus is particularly bright tonight." In any case, this stock consists of sentences and not of propositions.

The basic objects in the agents' *plans* are atomic actions and observations which may be active (one looks for something) or passive (one happens to see something). These are supplemented by the operations of concatenation (sequencing), *if then else*, and *while do*, where the tests in the *if then else* and *while do* are on *sentences*. There may also be recursive calls to the procedure: *find out if the sentence ϕ or its negation is derivable within the limits of my current resources, from my current stock of beliefs*. Thus if i's plan has currently a test on ϕ, then, to be sure, the stock of sentences will be consulted to see if ϕ or its negation is in the stock. But there may also be a recursive call to a procedure for deciding ϕ. If someone asks "Do you

[17]In Bicchieri (1997) suggests that co-operation also comes about as a result of such a learning process. Such suggestions have of course also been made by many others. Since we are only considering the one-agent case here, we shall not go into this issue any further. See, however, our Parikh (1991).

[18]It may seem to the reader as if I am endorsing a representational theory after all, but not so. First, the stock may not literally exist, but may simply refer to those sentences which the agent assents to *quickly*. Secondly, the agent's beliefs need not be restricted to this stock – just as earlier, the bookseller Jill was not restricted to selling sets of books which were in her store *as a set*.

know the time?", we do not usually say, "I don't", but look at our watches. Thus consulting our stock of sentences is typically only the first step in deciding if some sentence or its negation can be derived with the resources we have.

This difference between sentences and propositions matters as we now show.[19]

It has been suggested in this context (e.g. by Stalnaker 1999) that such issues can be addressed by using the notion of fine grain. By this account, if I understand it correctly, logical equivalence is too coarse a grain and that a finer grain may be needed. So if two sentences are in the same fine grain and an agent knows or believes one, then the agent will also know or believe the other. But if we try to flesh out this metaphor then we can see that it is not going to work.

For instance if $G(\phi, \psi)$ means that ϕ and ψ are in the same grain, then G will be an equivalence relation. But surely we cannot have transitivity in reality, because we could have a sequence $\phi_1,\ldots,\phi_n$ of sentences, any two successive ones of which are easily seen to be equivalent, whereas it is quite hard to see the equivalence of ϕ_1 and ϕ_n.

Moreover, "being in the same grain" sounds interpersonal. If two molecules are in the same rice grain for you, then they are also in the same fine grain for me – it is just a fact of the matter. But in reality people differ greatly in their ability to perceive logical equivalences.

Thus suppose some set theorist thinks of some new axiom ϕ and wonders if ϕ implies the continuum hypothesis, call it ψ. The set theorist may find it quite easy to decide this question even if we see no resemblance between ϕ and ψ. And if he does not find it easy, he may look in the literature or ask another set theorist, processes which cannot easily be built into a formal theory. And they should not be! For if they were easy to build into a formal theory, then the representation is almost certain to be wrong.

Or a chess champion may be able to see 20 moves (40 half-moves) ahead in the end game, but you and I cannot. And he too usually cannot do this during the middle game. Thus context, habit and expertise matter a lot.

Suppose for instance that Lois Lane has invited Clark Kent to dinner but he has not said yes or no. So she forms the plan, *While I do not have a definite answer one way or another,* if *I see Clark Kent, I will ask him if he is coming to dinner.* Here *seeing Clark Kent* is understood to consist of an observation followed by the addition of the sentence "I am seeing Clark Kent" to her stock.

Suppose now that she sees Superman standing on her balcony. She will *not* ask him if he is coming to dinner as the sentence "I am seeing Clark Kent" will not be in

[19]We might compare entertaining a proposition as a bit like entering a building. If Ann and Bob enter the same building through different doors, they need not be in the same spot, and indeed they might never meet. But what makes it the same building is that they *could* meet without going out into the street. Thus if Ann and Bob are apt to assent to sentences s, s' respectively where *we* know that s, s' are equivalent; then it need not follow that there is a sentence they share. But they *could* through a purely deductive process, and without appealing to any additional facts, be brought to a situation where Ann assents to s' and Bob to s (unless one of them withdraws a belief, which may also happen).

her stock of sentences. And this is the sense in which she does *not* know that when she is seeing Superman, she is also seeing Clark Kent. If she *suspects* that Clark Kent is Superman, then it may happen that her recursive call to the procedure "decide if I am seeing Clark Kent" will take the form of the question, "Are you by any chance Clark Kent, and if so, are you coming to dinner?" addressed to Superman.

Have we given up too much by using sentences and not propositions as the objects of i-belief? Suppose her plan is, "If I see Jack and Jill, I will ask them to dinner" but she sees Jill first and then Jack so that she adds the sentence "Jill and Jack are in front of me" to her stock. Will this create a problem? Not so, because if a sentence ϕ is in her stock, ϕ easily implies ψ, and she needs to know the value of ψ so she can choose, then the program *find out about* ψ which she calls will probably find the sentence she needs. If the program terminates without yielding an answer she may well have a default action which she deems safest.

So here we make use of the fact that Lois does have a reasonable amount of intelligence. Even if she does not explicitly add some sentence ψ to her stock, when she comes to a point where the choice of action depends on ψ, she *will* ask if ψ or its negation is derivable from her present stock of sentences, possibly supplemented by some actions which add to this stock.

Definition 3. If an agent a comes to a point in her plan where her appropriate action is *If* ϕ *then do* α *else do* β, and she does α, then we will say that she *i-believes* ϕ. If, moreover, ϕ is true, and we believe that in a similar context she would judge it to be true only if it *is* true, then (within the context of this plan) we will say that she *i-knows* ϕ.

A common example of such a plan is the plan to answer a question correctly. Thus if an agent is asked "Is ϕ true?", the agent will typically call the procedure "decide if ϕ is true", and then answer "yes", "no", or "I don't know" in the appropriate cases.

Now note that if an agent means to deceive, then the same procedure will be called, but the answers given will be the opposite of the ones indicated by the procedure. But if we ourselves know that the agent's plan is to deceive, then we can clearly take the statement "ϕ is false" to indicate that the agent believes ϕ.

We no longer have the law that if the agent i-knows ϕ and ϕ implies ψ then the agent of necessity i-knows ψ. But if the agent has the resources to decide ϕ and the proof of ψ from ϕ is easy, then she might well also know ψ. But her notion of "easy" may be different from ours, and how much effort she devotes to this task will depend on her mood, how much energy she has, etc. Dennett (1985), who makes a related point, does not seem to make the distinction between linguistic and non-linguistic belief which I am making here.

For instance if a customer sits down on a bar stool and the bartender sees him, we do not need to ask, "Does the bartender know he wants a drink?" Of course he knows. But suppose a woman suffering from a persistent cough calls her husband and says, "I got an appointment with the doctor for 2:30 PM", she may later think, "I wonder if he realized that I cannot pick up our daughter at 3". When i knows ϕ and ψ is deducible from ϕ, then whether we assume that i knows ψ will depend less on some objective distance between ϕ and ψ than on what we know or can assume about i's habits.

Note now that we often tend to assign knowledge and beliefs to agents even when they are not in the midst of carrying out a plan. Even when we are brushing our teeth we are regarded as knowing that the earth goes around the sun. This can be explained by a continuity assumption. Normally when we *are asked* if the earth goes around the sun or vice versa, we say that the former is the case. An agent is then justified in assuming that even in between different occasions of being so asked, if we *were* asked, we would give the same answer. This assumption, which is usually valid, is what accounts for such attributions of belief.

Conclusion

We have tried to give an account of belief which starts with behavior rather than with some ad hoc logical system. There are many things to be worked out. For instance, why do we perform deductions at all? They must benefit us in some way, but a detailed logical account of how systems can evolve towards better logical acumen is yet to be developed. Yet another question to ask is about partial beliefs. In some circumstances these partial beliefs will be correlated with a probability function, but again, a concordance with the Kolmogorov axioms will be a norm which is not always observed in practice.

Acknowledgements We thank Sergei Artemov, Can Başkent, Samir Chopra, Horacio Arló Costa, Juliet Floyd, Haim Gaifman, Isaac Levi, Mike Levin, Larry Moss, Eric Pacuit, Catherine Wilson, and Andreas Witzel for comments. The information about chess came from Danny Kopec. This research was supported by a grant from the PSC-CUNY faculty research assistance program. Earlier versions of this paper were given at *TARK-05*, *ESSLLI-2006*, at the *Jean Nicod Institute*, at a seminar in the philosophy department at Bristol University, and at the Philosophy Colloquium at the City University Graduate Center. Some of the research for this paper was done when the author was visiting Boston University and the Netherlands Institute for Advanced Study. A very preliminary version of some of the ideas was presented at Amsterdam, and published as Parikh (2001). This research was partially supported by grants from the PSC-CUNY program at the City university of New York.

References

Alchourron, C., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic, 50*, 510–530.

Artemov, S., & Nogina, E. (2005). On epistemic logics with justifications. In R. Meyden (Ed.), *Theoretical aspecits of rationality and knowledge* (pp. 279–294). Singapore: University of Singapore press.

Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics, 4*, 1236–1239.

van Benthem, J. (1976). *Modal correspondence theory*. Doctoral dissertation, University of Amsterdam.

Bicchieri, C. (1997). Learning to co-operate. In C. Biccheri, R. C. Jeffrey, & B. Skyrms (Eds.), *The dynamics of norms* (pp. 17–46). Cambridge: Cambridge University Press.

Brandom, R. (1994). Unsuccessful seminatics. *Analysis, 54*, 175–178.
Davidson, D. (1982). Rational animals. *Dialectica, 36*, 318–327.
Dennett, D. (1985). *Brainstorms* (2nd ed.). MIT, cf. page 11. Singapore.
Dennett, D. (1996). *The Intentional Stance (6th printing)*, Cambridge, Massachusetts: The MIT Press.
Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. (1995). *Reasoning about knowledge*. Cambridge: MIT.
de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. *Annales de l'Institut Henri Poincare, 7*. (Trans.: Kyburg, H. (1980). *Studies in subjective probability*, Kyburg and Smokler (Eds.), (pp. 53–118). Krieger Publishing.)
Fitting, M. (2004). A logic of explicit knowledge. *Logica Yearbook*, 11–22.
Gaifman, H. (2004). Reasoning with limited resources and assigning probabilities to arithmetical statements. *Synthese, 140*, 97–119.
Gettier, E. (1963). Is justified true belief knowledge? *Analysis, 23*, 121–123.
Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103*, 592–596.
Hayek, F. A. (1948). *Individualism and economic order* (See especially chapters II and IV). Chicago: University of Chicago Press.
Hintikka, J. (2004). A fallacious fallacy? *Synthese, 140*, 25–35.
Hume, D. (1988). *A treatise of human nature*, (L. A. Selby-Brigge (Ed.), pp. 176–179). New York: Oxford University Press.
Kahneman, D. (2002). *Maps of Bounded Rationality*, Nobel prize lecture.
Kripke, S. (1979). A puzzle about belief. In A. Margalit (Ed.), *Meaning and use*. Dordrecht/Boston: Reidel
Levi, I. (1997). Rationality and commitment. In *The covenant of reason*. Cambridge/New York: Cambridge University Press.
Levi, I. (2004). Jaakko Hintikka. *Synthese, 140*, 37–41.
Marcus, R. (1990). Some revisionary proposals about belief and believing. *Philosophy and Phenomenological Research, 50*, 133–153.
Marcus, R. (1995). The anti-naturalism of some language centere accounts of belief. *Dialectica, 49*, 112–129.
Millikan, R. (2006). Styles of rationality. In M. Nudds & S. Hurley (Eds.), *Rationality in animals* (pp. 117–126). Oxford: Oxford University Press.
Milner, R. (1989). *Communication and concurrency*. New York: Prentice Hall.
Moses, Y. (1988). Resource bounded knowledge. In M. Vardi (Ed.), *Proceedings of the 2nd Conference on Theorectical Aspects of Reasoning About Knowledge* (pp. 261–276). San Francisco: Morgan Kaufmann.
Parikh, R. (1987). Knowledge and the problem of logical omniscience. In *International symposium on methodology for intelligent systems (ISMIS- 87)*, Charlotte (pp. 432–439). North Holland.
Parikh, R. (1991). Finite and infinite dialogues. In Moschovakis (Ed.) *Proceedings of a Workshop on Logic from Computer Science* (pp. 481–498). Berlin: Springer/MSRI Publications.
Parikh, R. (2001). Propositions, propositional attitudes and belief revision. In K. Segerberg, M. Zakharyaschev, M. de Rijke, & H. Wansing (Eds.), *Advances in modal logic* (Vol. 2). Stanford: CSLI Publications.
Parikh, R. (1995). Logical omniscience. In Leivant (Ed.), *Logic and computational complexity* (Lecture notes in computer science, Vol. 960, pp. 22–29). Berlin/New York: Springer.
Parikh, R., & Ramanujam, R. (2003). A knowledge based semantics of messages. *Journal of Logic, Language and Information, 12*, 453–467.
Park, D. (1981). Concurrency and automata on infinite sequences. In P. Deussen (Ed.), *Proceedings of the 5th GI-Conference Karlsruhe*. Berlin/Heidelberg: Springer.
Pepperberg, I. (2004). Talking with Alex: Logic and speech in parrots; exploring intelligence. *Scientific American Mind,* August 2004.
Plato. *Meno*. (380 BC). (trans. Benjamin Jowett). Available online at http://classics.mit.edu//Plato/meno.html

Ramsey, F. P. (1990). Facts and propositions. In D. H. Mellor (Ed.), *Philosophical papers* (pp. 34–51). Cambridge/New York: Cambridge University Press.
Ramsey, F. P. (1931). Truth and probability. In *The foundations of mathematics* (pp. 156–198). London: Routledge and Kegan Paul.
Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Nous, 36:2*, 249–275.
Schwitzgebel, E. (2006, Fall). Belief. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Fall 2006 ed.). http://plato.stanford.edu/archives/fall2006/entries/belief/.
Searle, J. (1994). Animal minds. In P. French & H. Wettstein (Eds.), *Philosophical naturalism, midwest studies in philosophy* (XIX, pp. 206–219). Notre Dame: University of Notre Dame Press.
Stalnaker, R. (1999). *Context and content*. Oxford/New York: Oxford University Press.
de Waal, F. (2005). *Our inner ape*. Penguin. Singapore.
Whyte, J. T. (1990). Success semantics. *Analysis, 50*, 149–157.
Wittgenstein, L. (1958). *Philosophical investigations*. New York: MacMillan.

Chapter 32
The Logic of Justification

Sergei Artemov

Introduction

The celebrated account of Knowledge as Justified True Belief commonly attributed to Plato (cf. Gettier 1963; Hendricks 2003) was widely accepted until 1963 when a paper by Edmund Gettier (1963) opened the door to a broad philosophical discussion of the subject (cf. Dretske 1971; Goldman 1967; Lehrer and Paxson 1969; Nozick 1981; Stalnaker 1996 and many others).

Meanwhile, commencing from seminal works (Hintikka 1962; von Wright 1951), the notions of Knowledge and Belief have acquired formalization by means of modal logic with atoms $\mathbf{K}F$ (*F is known*) and $\mathbf{B}F$ (*F is believed*). Within this approach, the following analysis was adopted: for a given agent,

$$F \text{ is known} \quad \sim \quad F \text{ holds in all epistemically possible situations.} \tag{32.1}$$

The resulting *Epistemic Logic* has been remarkably successful in terms of developing a rich mathematical theory and applications (cf. Fagin et al. 1995; Meyer and van der Hoek 1995, and other sources). However, the notion of justification, which has been an essential component of epistemic studies, was conspicuously absent in the mathematical models of knowledge within the epistemic logic framework. This deficiency is displayed most prominently, in the *Logical Omniscience* defect of the modal logic of knowledge (cf. Fagin and Halpern 1985, 1988; Hintikka 1975;

This work has been supported by NSF grant 0830450, CUNY Collaborative Incentive Research Grant CIRG1424, and PSC CUNY Research Grant PSCREG-39-721.

S. Artemov (✉)
Graduate Center CUNY, 365 Fifth Avenue, New York, NY 10016, USA
e-mail: sartemov@gc.cuny.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_32

Moses 1988; Parikh 1987). In the provability domain, the absence of an adequate description of the logic of justifications (here mathematical proofs) remained an impediment to both formalizing the Brouwer-Heyting-Kolmogorov semantics of proofs and providing a long-anticipated exact provability semantics for Gödel's provability logic S4 and intuitionistic logic (Artemov 1999, 2001, 2007; van Dalen 1986). This lack of a justification component has, perhaps, contributed to a certain gap between epistemic logic and mainstream epistemology (Hendricks 2003, 2005). We would like to think that Justification Logic is a step towards filling this void.

The contribution of this paper to epistemology can be briefly summarized as follows.

> *We describe basic logical principles for justifications and relate them to both mainstream and formal epistemology. The result is a long-anticipated mathematical notion of justification, making epistemic logic more expressive. We now have the capacity to reason about justifications, simple and compound. We can compare different pieces of evidence pertaining to the same fact. We can measure the complexity of justifications, which leads to a coherent theory of logical omniscience. Justification Logic provides a novel, evidence-based mechanism of truth-tracking which seems to be a key ingredient of the analysis of knowledge. Finally, Justification Logic furnishes a new, evidence-based foundation for the logic of knowledge, according to which*

$$F \text{ is known} \quad \sim \quad F \text{ has an adequate justification}. \tag{32.2}$$

There are several natural interpretations of Justification Logic. Justification assertions of the format $t : F$ read generically as

$$t \text{ is a justification of } F. \tag{32.3}$$

There is also a more strict 'justificationist' reading in which $t : F$ is understood as

$$t \text{ is accepted by agent as a justification of } F. \tag{32.4}$$

The language and tools of Justification Logic accommodate both readings of $t : F$. Moreover, Justification Logic is general enough to incorporate other semantics that are not necessarily terminologically related to justifications or proofs. For example, $t : F$ can be read as

$$t \text{ is a sufficient resource for } F. \tag{32.5}$$

Tudor Protopopescu suggests that $t : F$ could also be assigned an externalist, non-justificationist reading, something like

$$F \text{ satisfies conditions } t. \tag{32.6}$$

In this setting, t would be something like a set of causes or counterfactuals. Such a reading would still maintain the distinction between partial and factive justifications, since t may not be all that is required for *belief that F* to count as *knowledge that F*.

Within Justification Logic, we do not directly analyze what it means for t to justify F beyond the format $t : F$, but rather attempt to characterize this relation axiomatically. This is similar to the way Boolean logic treats its connectives, say, disjunction: it does not analyze the formula $p \vee q$ but rather assumes certain logical axioms and truth tables about this formula.

There are several design decisions made for this installment of Justification Logic.

1. We decide to limit our attention, at this stage, to *propositional* and *quantifier-free* systems of Justification Logic, and leave quantified systems for further study.
2. We build our systems on the simplest base: *classical Boolean logic*, though we are completely aware that there are much more elaborate logical models, e.g., intuitionistic and substructural logics, conditionals, relevance logics, and logics of counterfactual reasoning, just to name a few. There are several good reasons for choosing the Boolean logic base here. At this stage, we are concerned first with *justifications*, which provide a sufficiently serious challenge on even the simplest Boolean base. Once this case is sorted out in a satisfactory way, we can move on to incorporating justifications into other logics. Second, the paradigmatic examples which we will consider (e.g., Goldman-Kripke and Gettier), can be handled with Boolean Justification Logic. Third, the core of Epistemic Logic consists of modal systems with a classical Boolean base (K, T, K4, S4, K45, KD45, S5, etc.). We provide each of them with a corresponding Justification Logic companion based on Boolean logic.
3. Within the Justification Logic framework, we treat both partial and factive justifications. This helps to capture the essence of discussion on these matters in epistemology, where justifications are not generally assumed to be factive.
4. In this paper, we consider the case of one agent only, although several multi-agent Justification Logic systems have already been developed (Artemov 2006; Artemov and Nogina 2005; Yavorskaya (Sidon) 2006).

Formal logical methods do not directly solve philosophical problems, but rather provide a tool for analyzing assumptions and to ensure that we draw correct conclusions. Our hope is that Justification Logic will do just that.

Preliminary Analysis of Principles Involved

In this section, we will survey the Logic of Proofs, Gettier's examples (Gettier 1963), and examine some classical post-Gettier sources to determine what logical principles in the given Justification Logic format (propositional Boolean logic with justification assertions $t : F$) may be extracted. As is usual with converting informally stated principles into formal ones, a certain amount of good will is required. This does not at all mean that the considerations adduced in Dretske (1971), Goldman (1967), Lehrer and Paxson (1969), Nozick (1981), and Stalnaker (1996) may be readily formulated in the Boolean Justification Logic. The aforementioned

papers are written in natural language, which is richer than any formal one; a more sophisticated formal language could probably provide a better account here, which we leave to future studies.

The Logic of Proofs

The Logic of Proofs LP was suggested by Gödel in (1995) and developed in full in Artemov (1995, 2001). LP gives a complete axiomatization of the notion of mathematical proof with natural operations 'application,' 'sum,' and 'proof checker.' We discuss these operations below in a more general epistemic setting.

In LP, justifications are represented by *proof polynomials*, which are terms built from *proof variables* $x, y, z, \ldots$ and *proof constants* $a, b, c, \ldots$ by means of two binary operations: *application* '$\cdot$' and *sum* (*union, choice*) '$+$,' and one unary operation *proof checker* '!'. The formulas of LP are those of propositional classical logic augmented by the formation rule: *if t is a proof polynomial and F a formula, then $t : F$ is again a formula.*

The Logic of Proofs LP contains the postulates of classical propositional logic and the rule of *Modus Ponens* along with

$s : (F \to G) \to (t : F \to (s \cdot t) : G)$ (*Application*)
$s : F \to (s + t) : F, \quad t : F \to (s + t) : F$ (*Sum*)
$t : F \to !t : (t : F)$ (*Proof Checker*)
$t : F \to F$ (*Reflection*).

Proof constants in LP represent 'atomic' proofs of axioms which are not analyzed any further. In addition to the usual logical properties, such as being closed under substitution and respecting the Deduction Theorem, LP enjoys the Internalization property:

If $\vdash F$, then there is a proof polynomial p such that $\vdash p : F$.

Gettier Examples

Gettier in Gettier (1963) described two situations, Case I and Case II, that were supposed to provide examples of justified true beliefs which should not be considered knowledge. In this paper we will focus on formalizing Case I, which proved to be more challenging. Case II can be easily formalized in a similar fashion.

Here is a shortened exposition of Case I from Gettier (1963).

Suppose that Smith and Jones have applied for a certain job. And suppose that Smith has strong evidence for the following conjunctive proposition:

(d) Jones is the man who will get the job, and Jones has ten coins in his pocket. Proposition (d) entails:

(e) The man who will get the job has ten coins in his pocket.

Let us suppose that Smith sees the entailment from (d) to (e), and accepts (e) on the grounds of (d), for which he has strong evidence. In this case, Smith is clearly justified in believing that (e) is true. But imagine, further, that unknown to Smith, he himself, not Jones, will get the job. And, also, unknown to Smith, he himself has ten coins in his pocket. Then, all of the following are true:

(1) (e) is true,
(2) Smith believes that (e) is true, and
(3) Smith is justified in believing that (e) is true.

But it is equally clear that Smith does not know that (e) is true. . ..

Gettier uses a version of the epistemic closure principle, closure of justification under logical consequence:

> *. . . if Smith is justified in believing P, . . . and Smith deduces Q from P . . ., then Smith is justified in believing Q.*

Here is its natural formalization:

> *Smith is justified in believing P* can be formalized as "for some t, $t : P$";
> *Smith deduces Q from P*—"there is a deduction of $P \to Q$ (available to Smith)";
> *Smith is justified in believing Q*—"$t : Q$ for some t."

Such a rule holds for the Logic of Proofs, as well as for all other Justification Logic systems considered in this paper. It is a combination of the Internalization Rule:

$$\textit{if} \vdash F, \textit{ then} \vdash s : F \textit{ for some } s \tag{32.7}$$

and the Application Axiom:

$$s : (P \to Q) \to (t : P \to (s \cdot t) : Q). \tag{32.8}$$

Indeed, suppose $t : P$ and there is a deduction of $P \to Q$. By the Internalization Rule, $s : (P \to Q)$ for some s. From the Application Axiom, by *Modus Ponens* twice, we get $(s \cdot t) : Q$.

Goldman's Reliabilism

Goldman in Goldman (1967) offered the 'fourth condition' to be added to the Justified True Belief definition of knowledge. According to Goldman (1967),

> *a subject's belief is justified only if the truth of a belief has caused the subject to have that belief (in the appropriate way), and for a justified true belief to count as knowledge, the subject must also be able to 'correctly reconstruct' (mentally) that causal chain.*

Goldman's principle makes it clear that a justified belief (in our language, a situation *t justifies F* for some *t*) for an agent occurs **only if** *F* is true, which provides the Factivity Axiom for 'knowledge-producing' justifications

$$t : F \to F \qquad \textit{(Factivity Axiom)}. \tag{32.9}$$

The Factivity Axiom is assumed for *factive justifications* (systems JT, LP, JT45 below) but not for general justification systems J, J4, J45, JD45.

With a certain amount of good will, we can assume that *the 'causal chain' leading from the truth of F to a justified belief that F* manifests itself in the Principle of Internalization which holds for many Justification Logic systems:

If F is valid, then one could construct a justification p such that $p : F$ *is valid.* (32.10)

Internalization is usually represented in an equivalent form (in the presence of the Completeness Theorem) as a meta-rule (32.7). The algorithm which builds a justified belief $p : F$ from a strong evidence (proof) of the validity of *F* seems to be an instance of Goldman's 'causal chain.'

Lehrer and Paxson's Indefeasibility Condition

Lehrer and Paxson in Lehrer and Paxson (1969) offered the following 'indefeasibility condition':

> *there is no further truth which, had the subject known it, would have defeated [subject's] present justification for the belief.*

The 'further truth' here could refer to a possible update of the subject's database, or some possible-worlds situation, etc.: these readings lie outside the scope of our language of Boolean Justification Logic. A natural reading of 'further truth' in our setting could be 'other postulate or assumption of the system,' which means a simple consistency property which vacuously holds for all Justification Logic systems considered here. Another plausible reading of 'further truth' could be 'further evidence,' and we assume this particular reading here. Since there is no temporal or update component in our language yet, 'any further evidence' could be understood for now as 'any other justification,' or just 'any justification.'

Furthermore, Lehrer and Paxson's condition seems to involve a negation of an existential quantifier over justifications 'there is no further truth . . . ,' or

there is no justification. . ..

However, within the classical logic tradition, we can read this as a universal quantifier over justifications followed by a negation

for any further evidence, it is not the case. . ..

Denoting 'present justification for the belief' as the assertion $s : F$, we reformulate Lehrer–Paxson's condition as

given $s : F$, for any evidence t, it is not the case that t would have defeated $s : F$.

The next step is to formalize 't does not defeat $s : F$.' This informal statement seems to suggest an implication

if $s : F$ holds, then the joint evidence of s and t, which we denote here as $s + t$, is also an evidence for F, i.e., $(s + t) : F$ holds.

Here is the resulting formal version of Lehrer–Paxson's condition: for any proposition F and any justifications s and t, the following holds

$$s : F \to (s + t) : F \qquad \textit{(Monotonicity Axiom).} \tag{32.11}$$

Further Assumptions

In order to build a formal account of justification, we will make some basic structural assumptions: *justifications are abstract objects which have structure, operations on justifications are potentially executable, agents do not lose or forget justifications, agents apply the laws of classical logic and accept their conclusions, etc.*

In the following, we consider both: **justifications**, which do not necessarily yield the truth of a belief, and **factive justifications**, which yield the truth of the belief.

Basic Principles and Systems

Application

The *Application* operation takes justifications s and t and produces a justification $s \cdot t$ such that if $s : (F \to G)$ and $t : F$, then $(s \cdot t) : G$. Symbolically,

$$s : (F \to G) \to (t : F \to (s \cdot t) : G). \tag{32.12}$$

This is a basic property of justifications assumed in combinatory logic and λ-calculi (cf. Troelstra and Schwichtenberg 1996), BHK-semantics (Troelstra and van Dalen 1988), Kleene realizability (Kleene 1945), the Logic of Proofs LP (Artemov 2001), etc. Application Principle (32.12) is related to the epistemological closure principle (cf., for example, Dretske 2005; Luper 2005) that one knows everything that one knows to be implied by what one knows. However, (32.12) does not rely on this closure principle, since (32.12) deals with a broader spectrum of justifications, not necessarily linked to knowledge.

Note that the epistemological closure principle which could be formalized using the knowledge modality **K** as

$$\mathbf{K}(F \to G) \to (\mathbf{K}F \to \mathbf{K}G), \tag{32.13}$$

smuggles the *logical omniscience* defect into modal epistemic logic. The latter does not have the capacity to measure how hard it is to attain knowledge (Fagin and Halpern 1985, 1988; Hintikka 1975; Moses 1988; Parikh 1987). Justification Logic provides natural means of escaping logical omniscience by keeping track of the size of justification terms Artemov and Kuznets (2006).

Monotonicity of Justification

The *Monotonicity* property of justification has been expressed by the operation *sum* '+,' which can be read from (32.11). If $s : F$, then whichever evidence t occurs, the combined evidence $s + t$ remains a justification for F. Operation '+' takes justifications s and t and produces $s + t$, which is a justification for everything justified by s or by t.

$$s : F \to (s + t) : F \quad \text{and} \quad s : F \to (t + s) : F.$$

A similar operation '+' is present in the Logic of Proofs LP, where the sum '$s + t$' can be interpreted as a concatenation of proofs s and t.

Correspondence Theorem 10 uses Monotonicity to connect Justification Logic with epistemic modal logic. However, it is an intriguing challenge to develop a theory of non-monotonic justifications which prompt belief revision. Some Justification Logic systems without Monotonicity have been studied in Artemov and Strassen (1993) and Krupski (2001, 2006).

Basic Justification Logic $\mathsf{J_0}$

Justification terms (polynomials) are built from justification variables $x, y, z, \ldots$ and justification constants $a, b, c, \ldots$ (with indices $i = 1, 2, 3 \ldots$ which we will be omitting whenever it is safe) by means of the operations **application** '$\cdot$' and **sum** '+'.[1] Constants denote atomic justifications which the system no longer analyzes; variables denote unspecified justifications.

Basic Logic of Justifications J_0:

[1]More elaborate models considered below in this paper also use additional operations on justifications, e.g., verifier '!' and negative verifier '?'.

A1. *Classical propositional axioms and rule Modus Ponens,*
A2. *Application Axiom* $s : (F \to G) \to (t : F \to (s \cdot t) : G)$,
A3. *Monotonicity Axiom* $s : F \to (s + t) : F$, $s : F \to (t + s) : F$,

J_0 is the logic of general (not necessarily factive) justifications for an absolutely skeptical agent for whom no formula is provably justified, i.e., J_0 does not derive $t : F$ for any t and F. Such an agent is, however, capable of making *relative justification conclusions* of the form

$$\textit{if } x : A, y : B, \ldots, z : C \textit{ hold, then } t : F.$$

J_0 is able, with this capacity, to adequately emulate other Justification Logic systems in its language.

Logical Awareness and Constant Specifications

The *Logical Awareness principle* states that logical axioms are justified ex officio: an agent accepts logical axioms (including the ones concerning justifications) as justified. As stated here, Logical Awareness is too restrictive and Justification Logic offers a flexible mechanism of Constant Specifications to represent all shades of logical awareness.

Justification Logic distinguishes between an assumption and a justified assumption. Constants are used to denote justifications of assumptions in situations when we don't analyze these justifications any further. Suppose we want to postulate that an axiom A is justified for a given agent. The way to say it in Justification Logic is to postulate

$$e_1 : A$$

for some evidence constant e_1 with index 1. Furthermore, if we want to postulate that this new principle $e_1 : A$ is also justified, we can postulate

$$e_2 : (e_1 : A)$$

for the similar constant e_2 with index 2, etc. Keeping track of indices is not necessary, but it is easy and helps in decision procedures (cf. Kuznets 2008). The set of all assumptions of this kind for a given logic is called a *Constant Specification*. Here is a formal definition.

A **Constant Specification** *CS* for a given logic $\mathcal{L}$ is a set of formulas

$$e_n : e_{n-1} : \ldots : e_1 : A \quad (n \geq 1),$$

where A is an axiom of $\mathcal{L}$, and $e_1, e_2, \ldots, e_n$ are similar constants with indices $1, 2, \ldots, n$. We also assume that CS contains all intermediate specifications, i.e., whenever $e_n : e_{n-1} : \ldots : e_1 : A$ is in CS, then $e_{n-1} : \ldots : e_1 : A$ is in CS too. In this paper, we will distinguish the following types of constant specifications:

- *empty*: $CS = \emptyset$. This corresponds to an absolutely skeptical agent (cf. a comment after axioms of J_0).
- *finite*: CS is a finite set of formulas. This is a representative case, since any specific derivation in Justification Logic concerns only finite sets of constants and constant specifications.
- *axiomatically appropriate*: for each axiom A there is a constant e_1 such that $e_1 : A$ is in CS, and if

$$e_n : e_{n-1} : \ldots : e_1 : A \in CS,$$

  then

$$e_{n+1} : e_n : e_{n-1} : \ldots : e_1 : A \in CS.$$

  Axiomatically appropriate CS's are necessary for ensuring the Internalization property.
- *total*: for each axiom A and **any** constants $e_1, e_2, \ldots, e_n$,

$$e_n : e_{n-1} : \ldots : e_1 : A \in CS.$$

  We are reserving the name TCS for the total constant specification (for a given logic). Naturally, the total constant specification is axiomatically appropriate.

Logic of Justifications with given Constant Specification

$$\mathsf{J}_{CS} = \mathsf{J}_0 + CS.$$

Logic of Justifications

$$\mathsf{J} = \mathsf{J}_0 + R4,$$

where R4 is the **Axiom Internalization Rule**:

For each axiom A and any constants $e_1, e_2, \ldots, e_n$, infer $e_n : e_{n-1} : \ldots : e_1 : A$.

Note that J_0 is $\mathsf{J}_\emptyset$, and J coincides with J_{TCS}. The latter reflects the idea of the unrestricted Logical Awareness for J. A similar principle appeared in the Logic of Proofs LP; it has also been anticipated in Goldman's Goldman (1967). Note that any **specific** derivation in J may be regarded as a derivation in J_{CS} for a corresponding **finite** constant specification CS, hence finite CS's constitute an important representative class of constant specifications.

Logical Awareness expressed by axiomatically appropriate constant specifications is an explicit incarnation of the Necessitation Rule in modal epistemic logic:

$$\vdash F \quad \Rightarrow \quad \vdash \mathbf{K}F \tag{32.14}$$

applied to axioms.

Let us consider some basic examples of derivations in J. In Examples 1 and 2, only constants of level 1 have been used; in such situations we skip indices completely.

Example 1. This example shows how to build a justification of a conjunction from justifications of the conjuncts. In the traditional modal language, this principle is formalized as

$$\Box A \wedge \Box B \to \Box(A \wedge B).$$

In J we express this idea in a more precise justification language.

1. $A \to (B \to (A \wedge B))$, *a propositional axiom*;
2. $c : [A \to (B \to (A \wedge B))]$, *from 1, by R4*;
3. $x : A \to (c \cdot x) : (B \to (A \wedge B))$, *from 2, by A2 and Modus Ponens*;
4. $x : A \to (y : B \to ((c \cdot x) \cdot y) : (A \wedge B))$, *from 3, by A2 and some propositional reasoning*;
5. $x : A \wedge y : B \to ((c \cdot x) \cdot y) : (A \wedge B)$, *from 5, by propositional reasoning*.

The derived formula 5 contains constant c, which was introduced in line 2, and the complete reading of the result of this derivation is

$$x : A \wedge y : B \to ((c \cdot x) \cdot y) : (A \wedge B), \textit{ given } c : [A \to (B \to (A \wedge B))].$$

Example 2. This example shows how to build a justification of a disjunction from justifications of either of the disjuncts. In the usual modal language this is represented by

$$\Box A \vee \Box B \to \Box(A \vee B).$$

Let us see how this would look in J.

1. $A \to (A \vee B)$, *by A1*;
2. $a : [A \to (A \vee B)]$, *from 1, by R4*;
3. $x : A \to (a \cdot x) : (A \vee B)$, *from 2, by A2 and Modus Ponens*;
4. $B \to (A \vee B)$, *by A1*;
5. $b : [B \to (A \vee B)]$, *from 4, by R4*;
6. $y : B \to (b \cdot y) : (A \vee B)$ *from 5, by A2 and Modus Ponens*;
7. $(a \cdot x) : (A \vee B) \to (a \cdot x + b \cdot y) : (A \vee B)$, *by A3*;
8. $(b \cdot y) : (A \vee B) \to (a \cdot x + b \cdot y) : (A \vee B)$, *by A3*;
9. $(x : A \vee y : B) \to (a \cdot x + b \cdot y) : (A \vee B)$ *from 3, 6, 7, 8, by propositional reasoning*.

The complete reading of the result of this derivation is

$$(x : A \vee y : B) \to (a \cdot x + b \cdot y) : (A \vee B), \textit{ given } a : [A \to (A \vee B)] \textit{ and } b : [B \to (A \vee B)].$$

Explicit mention of Constant Specifications of Justification Logic systems is normally used when semantic issues are concerned: e.g., arithmetical, symbolic, and epistemic semantics. To define the truth value of a formula under a given interpretation, one should be given a specification of constants involved.

For each constant specification CS, J_{CS} enjoys the Deduction Theorem, because J_0 contains propositional axioms and *Modus Ponens* as the only rule of inference.

Theorem 3. *For each axiomatically appropriate constant specification CS,* J_{CS} *enjoys Internalization:*

$$\textit{If} \vdash F, \textit{ then} \vdash p : F \textit{ for some justification term } p.$$

Proof. Induction on derivation length. Suppose $\vdash F$. If F is an axiom, then, since CS is axiomatically appropriate, there is a constant e such that $e : F$ is in CS, hence an axiom of J_{CS}. If F is in CS, then, since CS is axiomatically appropriate, $e : F$ is in CS for some constant e. If F is obtained by *Modus Ponens* from $X \to F$ and X, then, by the Induction Hypothesis, $\vdash s : (X \to F)$ and $\vdash t : X$ for some s, t. By the Application Axiom, $\vdash (s \cdot t) : F$. Note that Internalization can require a growth of constant specification sets; if $\vdash F$ with a Constant Specification CS, then the proof of $p : F$ may need some Constant Specification CS' which is different from CS.

Q.E.D.

Red Barn Example and Tracking Justifications

We begin illustrating new capabilities of Justification Logic with a paradigmatic Red Barn Example which Kripke developed in 1980 in objection to Nozick's account of knowledge (cf. article *The Epistemic Closure Principle* in Stanford Encyclopedia of Philosophy (Luper 2005), from which we borrow the formulation, with some editing for brevity).

> *Suppose I am driving through a neighborhood in which, unbeknownst to me, papier-mâché barns are scattered, and I see that the object in front of me is a barn. Because I have barn-before-me percepts, I believe that the object in front of me is a barn. Our intuitions suggest that I fail to know barn. But now suppose that the neighborhood has no fake red barns, and I also notice that the object in front of me is red, so I know a red barn is there. This juxtaposition, being a red barn, which I know, entails there being a barn, which I do not, "is an embarrassment".*[2]

[2] Dretske (2005).

We proceed in the spirit of the Red Barn Example and consider it a general test for theories that explain knowledge. What we want is a way to represent what is going on here which maintains epistemic closure,

one knows everything that one knows to be implied by what one knows, (32.15)

but also preserves the problems the example was intended to illustrate.

We present plausible formal analysis of the Red Barn Example in epistemic modal logic (sections "Red Barn in Modal Logic of Belief" and "Red Barn in Modal Logic of Knowledge") and in Justification Logic (sections "Red Barn in Justification Logic of Belief" and "Red Barn in Justification Logic of Knowledge"). We will see that epistemic modal logic is capable only of telling us that there is a problem, whereas Justification Logic helps to analyse what has gone wrong. We see that closure holds as it is supposed to, and we see that if we keep track of justifications we can analyse why we had a problem.

Red Barn in Modal Logic of Belief

In our first formalization, the logical derivation will be made in epistemic modal logic with 'my belief' modality $\Box$. We then interpret some of the occurrences of $\Box$ as 'knowledge' according to the problem's description. We will not try to capture the whole scenario formally; to make our point, it suffices to formalize and verify its "entailment" part. Let

- B be 'the object in front of me is a barn,'
- R be 'the object in front of me is red,'
- $\Box$ be 'my belief' modality.

The formulation considers observations 'I see a barn' and 'I see a red barn,' and claims logical dependencies between them. The following is a natural formalization of these assumptions in the epistemic modal logic of belief:

1. $\Box B$, 'I believe that the object in front of me is a barn';
2. $\Box(B \wedge R)$, 'I believe that the object in front of me is a red barn.'

At the metalevel, we assume that 2 is knowledge, whereas 1 is not knowledge by the problem's description. So, we could add factivity of 2, $\Box(B \wedge R) \rightarrow (B \wedge R)$, to the formal description, but this would not matter for our conclusions. We note that indeed 1 logically follows from 2 in the modal logic of belief K:

3. $(B \wedge R) \rightarrow B$, *logical axiom*;
4. $\Box[(B \wedge R) \rightarrow B]$, *from 3, by Necessitation. As a logical truth, this is a case of knowledge too*;
5. $\Box(B \wedge R) \rightarrow \Box B$, *from 4, by modal logic*.

Within this formalization, it appears that Closure Principle (32.15) is violated: $\Box(B\wedge R)$ is knowledge by the problem's description, $\Box[(B\wedge R)\to B]$ is knowledge as a simple logical axiom, whereas $\Box B$ is not knowledge.

Red Barn in Modal Logic of Knowledge

Now we will use epistemic modal logic with 'my knowledge' modality **K**. Here is a straightforward formalization of Red Barn Example assumptions:

1. $\neg\mathbf{K}B$, 'I do not know that the object in front of me is a barn';
2. $\mathbf{K}(B\wedge R)$, 'I know that the object in front of me is a red barn.'

It is easy to see that these assumptions are inconsistent in the modal logic of knowledge. Indeed,

3. $\mathbf{K}(B\wedge R)\to(\mathbf{K}B\wedge\mathbf{K}R)$, *by normal modal logic*;
4. $\mathbf{K}B\wedge\mathbf{K}R$, *from 2 and 3, by Modus Ponens*;
5. $\mathbf{K}B$, *from 4, by propositional logic*.

Lines 1 and 5 formally contradict each other.

Modal logic of knowledge does not seem to apply here.

Red Barn in Justification Logic of Belief

Justification Logic seems to provide a more fine-grained analysis of the Red Barn Example. We naturally refine assumptions by introducing individual justifications u for belief that B, and v for belief that $B\wedge R$. The set of assumptions in the Justification Logic is

1. $u:B$, 'u is the reason to believe that the object in front of me is a barn';
2. $v:(B\wedge R)$, 'v is the reason to believe that the object in front of me is a red barn.' *On the metalevel, the description states that this is a case of knowledge, not merely a belief.*

Again, we can add the factivity condition for 2, $v:(B\wedge R)\to(B\wedge R)$, but this does not change the analysis here. Let us try to reconstruct the reasoning of the agent in J:

3. $(B\wedge R)\to B$, *logical axiom*;
4. $a:[(B\wedge R)\to B]$, *from 3, by Axiom Internalization. This is also knowledge, as before*;
5. $v:(B\wedge R)\to(a\cdot v):B$, *from 4, by Application and Modus Ponens*;
6. $(a\cdot v):B$, *from 2 and 5, by Modus Ponens*.

Closure holds! Instead of deriving 1 from 2 as in section "Red Barn in Modal Logic of Belief", we have obtained a correct conclusion that $(a \cdot v) : B$, i.e., 'I know B for reason $a \cdot v$,' which seems to be **different** from u: the latter is the result of a perceptual observation, whereas the former is the result of logical reasoning. In particular, we cannot conclude that 2, $v : (B \wedge R)$, entails 1, $u : B$; moreover, with some basic model theory of J in section "Basic Epistemic Semantics", we can show that 2 **does not entail** 1. Hence, after observing a red façade, I indeed know B, but this knowledge does not come from 1, which remains a case of belief rather than of knowledge.

Red Barn in Justification Logic of Knowledge

Within this formalization, $t : F$ is interpreted as

'I know F for reason t.'

As in section "Red Barn in Modal Logic of Knowledge", we assume

1. $\neg u : B$, 'u is not a sufficient reason to know that the object is a barn';
2. $v : (B \wedge R)$, 'v is a sufficient reason to know that the object is a red barn.'

This is a perfectly consistent set of assumptions in the logic of factive justifications

$$\mathsf{J} + \textit{Factivity Principle } (t : F \rightarrow F).$$

As in section "Red Barn in Justification Logic of Belief", we can derive $(a \cdot v) : B$ where $a : [(B \wedge R) \rightarrow B]$, but this does not lead to a contradiction. Claims $\neg u : B$ and $(a \cdot v) : B$ naturally co-exist. They refer to different justifications u and $a \cdot v$ of the same fact B; one of them insufficient and the other quite sufficient for my knowledge that B.

It appears that in sections "Red Barn in Justification Logic of Belief" and "Red Barn in Justification Logic of Knowledge", Justification Logic represents the structure of the argument made by Kripke in his Red Barn Example, and which was not captured by traditional epistemic modal tools. The Justification Logic formalization represents what seems to be happening in such a case; we can maintain closure of knowledge under logical entailment, even though 'barn' is not perceptually known.

In all fairness to modal tools, we could imagine a formalization of the Red Barn example in a sort of bi-modal language with distinct modalities for knowledge and belief. However, it seems that such a resolution will, intellectually, involve repeating Justification Logic arguments in a way that obscures, rather than reveals, the truth. Such a bi-modal formalization would distinguish $u : B$ from $(a \cdot v) : B$ not because they have different reasons (which reflects the true epistemic structure of the problem), but rather because the former is labelled 'belief' and

the latter 'knowledge.' But what if we need to keep track of a larger number of different unrelated reasons? By introducing a number of distinct modalities and then imposing various assumptions governing the inter-relationships between these modalities, one would essentially end up with a reformulation of the language of Justification Logic itself (with distinct terms replaced by distinct modalities). This suggests that there may not really be a 'halfway point' between the modal language and the language of Justification Logic, at least inasmuch as one tries to capture the essential structure of examples involving the deductive failure of knowledge (e.g., Kripke's Red Barn example). Accordingly, one is either stuck with modal logic and its inferior account of these examples or else moves to Justification Logic and its superior account of these examples. This move can either come about by taking a multi-modal language and imposing inter-dependencies on different modals–ending up with something essentially equivalent to the language of Justification Logic–or else one can use the language of Justification Logic from the start. Either way, all there is to move to is Justification Logic.

Basic Epistemic Semantics

The standard epistemic semantics for J has been provided by the proper adaptation of Kripke-Fitting models Fitting (2005) and Mkrtychev models Mkrtychev (1997).

A Kripke-Fitting **J-model** $\mathcal{M} = (W, R, \mathcal{E}, \Vdash)$ is a Kripke model $(W, R, \Vdash)$ enriched with an **admissible evidence function** $\mathcal{E}$ such that $\mathcal{E}(t, F) \subseteq W$ for any justification t and formula F. Informally, $\mathcal{E}(t, F)$ specifies the set of possible worlds where t is considered admissible evidence for F. The intended use of $\mathcal{E}$ is in the truth definition for justification assertions:

$u \Vdash t : F$ *if and only if*

1. *F holds for all possible situations, that is, $v \Vdash F$ for all v such that uRv;*
2. *t is an admissible evidence for F at u, that is, $u \in \mathcal{E}(t, F)$.*

An admissible evidence function $\mathcal{E}$ must satisfy the closure conditions with respect to operations '$\cdot$' and '$+$':

- *Application*: $\mathcal{E}(s, F \to G) \cap \mathcal{E}(t, F) \subseteq \mathcal{E}(s \cdot t, G)$. This condition states that whenever s is an admissible evidence for $F \to G$ and t is an admissible evidence for F, their 'product,' $s \cdot t$, is an admissible evidence for G.
- *Sum*: $\mathcal{E}(s, F) \cup \mathcal{E}(t, F) \subseteq \mathcal{E}(s + t, F)$. This condition guarantees that $s + t$ is an admissible evidence for F whenever either s is admissible for F or t is admissible for F.

These are natural conditions to place on $\mathcal{E}$ because they are necessary for making basic axioms of Application and Monotonicity valid.

We say that $\mathcal{E}(t, F)$ *holds at a given world* u if $u \in \mathcal{E}(t, F)$.

Given a model $\mathcal{M} = (W, R, \mathcal{E}, \Vdash)$, the forcing relation $\Vdash$ is extended from sentence variables to all formulas as follows: for each $u \in W$,

1. $\Vdash$ respects Boolean connectives at each world ($u \Vdash F \wedge G$ iff $u \Vdash F$ and $u \Vdash G$; $u \Vdash \neg F$ iff $u \nVdash F$, etc.);
2. $u \Vdash t : F$ iff $u \in \mathcal{E}(t, F)$ and $v \Vdash F$ for every $v \in W$ with uRv.

Note that an admissible evidence function $\mathcal{E}$ may be regarded as a Fagin-Halpern awareness function (Fagin et al. 1995) equipped with the structure of justifications.

A model $\mathcal{M} = (W, R, \mathcal{E}, \Vdash)$ *respects a Constant Specification CS at* $u \in W$ if $u \in \mathcal{E}(c, A)$ for all formulas $c : A$ from *CS*. Furthermore, $\mathcal{M} = (W, R, \mathcal{E}, \Vdash)$ *respects a Constant Specification CS* if $\mathcal{M}$ respects *CS* at each $u \in W$.

Theorem 4. *For any Constant Specification CS,* J_{CS} *is sound and complete for the class of all Kripke-Fitting models respecting CS.*

Proof. Fix a Constant Specification *CS* and consider J_{CS}.

Soundness is straightforward. Induction on derivations in J_{CS}. Let us check the axioms.

Application. Suppose $u \Vdash s : (F \to G)$ and $u \Vdash t : F$. Then, by the definition of forcing, $u \in \mathcal{E}(s, F \to G)$ and $u \in \mathcal{E}(t, F)$, hence, by the closure condition for $\mathcal{E}$, $u \in \mathcal{E}(s \cdot t, G)$. Moreover, for each v such that uRv, $v \Vdash F \to G$ and $v \Vdash F$, hence $v \Vdash G$. Thus $u \Vdash (s \cdot t) : G$ and $u \Vdash s : (F \to G) \to (t : F \to (s \cdot t) : G)$.

Sum. Suppose $u \Vdash t : F$. Then $u \in \mathcal{E}(t, F)$, hence, by the closure condition for $\mathcal{E}$, $u \in \mathcal{E}(s + t, F)$. In addition, $v \Vdash F$ for each v such that uRv, hence $u \Vdash (s + t) : F$. Thus $u \Vdash t : F \to (s + t) : F$.

Axioms from *CS* hold at each world, since the models respect *CS*. The Induction Step corresponds to the use of *Modus Ponens*, which is clearly a sound rule here.

To establish completeness, we use standard canonical model construction. The canonical model $\mathcal{M} = (W, R, \mathcal{E}, \Vdash)$ for J_{CS} is defined as follows:

- W is the set of all maximal consistent sets in J_{CS}. Following an established tradition, we denote elements of W as Γ, Δ, etc.;
- $\Gamma R \Delta$ iff $\Gamma^{\sharp} \subseteq \Delta$, where $\Gamma^{\sharp} = \{F \mid t : F \in \Gamma \text{ for some } t\}$;
- $\mathcal{E}(s, F) = \{\Gamma \in W \mid s : F \in \Gamma\}$;
- $\Gamma \Vdash p$ iff $p \in \Gamma$.

The Truth Lemma claims that for all F's,

$$\Gamma \Vdash F \quad \textit{if and only if} \quad F \in \Gamma.$$

This is established by standard induction on the complexity of F. The atomic cases are covered by the definition of '$\Vdash$.' The Boolean induction steps are standard. Consider the case when F is $t : G$ for some t and G.

If $t : G \in \Gamma$, then $G \in \Delta$ for all Δ such that $\Gamma R \Delta$ by the definition of R. By the Induction Hypothesis, $\Delta \Vdash G$. In addition, $\Gamma \in \mathcal{E}(t, G)$ by the definition of $\mathcal{E}$. Hence $\Gamma \Vdash t : G$, i.e., $\Gamma \Vdash F$.

If $t : G \notin \Gamma$, then $\Gamma \notin \mathcal{E}(t, G)$, i.e., $\Gamma \nVdash t : G$ and $\Gamma \nVdash F$.

Furthermore, $\mathcal{M}$ respects *CS* at each node. Indeed, by the construction of $\mathcal{M}$, $CS \subseteq \Gamma$ for each $\Gamma \in W$. By the Truth Lemma, $\Gamma \Vdash c : A$ for each $c : A \in CS$.

The conclusion of the proof of Theorem 4 is standard. Let F be not derivable in J_{CS}. Then the set $\{\neg F\}$ is consistent. Using the standard saturation construction (Fagin et al. 1995; Meyer and van der Hoek 1995), extend $\{\neg F\}$ to a maximal consistent set Γ. By consistency, $F \notin \Gamma$. By the Truth Lemma, $\Gamma \nVdash F$. Q.E.D.

There are several features of the canonical model which could be included into the formulation of the Completeness Theorem to make it stronger.

Strong Evidence. We can show that the canonical model considered in this proof satisfies the Strong Evidence property

$$\Gamma \in \mathcal{E}(t, F) \quad \textit{implies} \quad \Gamma \Vdash t : F.$$

Indeed, let $\Gamma \in \mathcal{E}(t, F)$. By the definition of $\mathcal{E}$, $t : F \in \Gamma$, hence $F \in \Gamma^\sharp$ and $F \in \Delta$ for each Δ such that $\Gamma R \Delta$. By the Truth Lemma, $\Delta \Vdash F$, hence $\Gamma \Vdash t : F$. In a model with the Strong Evidence property there are no void or irrelevant justifications; if t is an admissible evidence for F, then t is a 'real evidence' for F, i.e., F holds at all possible worlds.

Fully Explanatory property for axiomatically appropriate Constant Specifications:

$$\textit{If } \Delta \Vdash F \textit{ for all } \Delta \textit{ such that } \Gamma R \Delta, \textit{ then } \Gamma \Vdash t : F \textit{ for some } t.$$

Note that for axiomatically appropriate constant specifications CS, the Internalization property holds: if G is provable in J_{CS}, then $t : G$ is also provable there for some term t. Here is the proof of the Fully Explanatory property for canonical models.[3] Suppose $\Gamma \nVdash t : F$ for any justification term t. Then the set $\Gamma^\sharp \cup \{\neg F\}$ is consistent. Indeed, otherwise for some $t_1 : X_1, t_2 : X_2, \ldots, t_n : X_n \in \Gamma$, $X_1 \to (X_2 \to \ldots \to (X_n \to F) \ldots)$ is provable. By Internalization, there is a justification s such that $s : (X_1 \to (X_2 \to \ldots \to (X_n \to F) \ldots))$ is also provable. By Application, $t_1 : X_1 \to (t_2 : X_2 \to \ldots \to (t_n : X_n \to (s \cdot t_1 \cdot t_2 \cdot \ldots \cdot t_n) : F) \ldots)$ is provable, hence $\Gamma \vdash t : F$ for $t = s \cdot t_1 \cdot t_2 \cdot \ldots \cdot t_n$. Therefore, $\Gamma \Vdash t : F$—a contradiction. Let Δ be a maximal consistent set extending $\Gamma^\sharp \cup \{\neg F\}$. By the definition of R, $\Gamma R \Delta$, by the Truth Lemma, $\Delta \nVdash F$, which contradicts the assumptions.

Mkrtychev semantics is a predecessor of Kripke-Fitting semantics (Mkrtychev 1997). *Mrktychev models* are Kripke-Fitting models with a single world, and the proof of Theorem 4 can be easily modified to establish completeness of J_{CS} with respect to Mkrtychev models.

Theorem 5. *For any Constant Specification CS, J_{CS} is sound and complete for the class of Mrktychev models respecting CS.*

Proof. Soundness follows immediately from Theorem 4. For completeness, define the canonical model as in Theorem 4 except for R, which should be taken empty.

[3] This proof for LP was offered by Fitting in Fitting (2005).

This assumption makes the condition '$\Delta \Vdash F$ for all Δ such that $\Gamma R \Delta$' vacuously true, and the forcing condition for justification assertions $\Gamma \Vdash t : F$ becomes equivalent to $\Gamma \in \mathcal{E}(t, F)$, i.e., $t : F \in \Gamma$. This simplification immediately verifies the Truth Lemma.

The conclusion of the proof of Theorem 5 is standard. Let F be not derivable in J_{CS}. Then the set $\{\neg F\}$ is consistent. Using the standard saturation construction, extend it to a maximal consistent set Γ containing $\neg F$. By consistency, $F \notin \Gamma$. By the Truth Lemma, $\Gamma \nVdash F$. The Mkrtychev model consisting of this particular Γ is the desired counter-model for F. The rest of the canonical model is irrelevant.

Q.E.D.

Note that Mkrtychev models built in Theorem 5 are not reflexive, and possess the Strong Evidence property. On the other hand, Mkrtychev models cannot be Fully Explanatory, since '$\Delta \Vdash F$ for all Δ such that $\Gamma R \Delta$' is vacuously true, but $\Gamma \Vdash t : F$ is not.

Theorem 5 shows that the information about Kripke structure in Kripke-Fitting models can be completely encoded by the admissible evidence function. Mkrtychev models play an important theoretical role in Justification Logic (Artemov 2008; Brezhnev and Kuznets 2006; Krupski 2006; Kuznets 2000; Milnikel 2007). On the other hand, as we will see in section "Formalization of Gettier Examples", Kripke-Fitting models can be useful as counter-models with desired properties since they take into account both epistemic Kripke structure and evidence structure. Speaking metaphorically, Kripke-Fitting models naturally reflect two reasons why a certain fact F can be unknown to an agent: F fails at some possible world or an agent does not have a sufficient evidence of F.

Another application area of Kripke-Fitting style models is Justification Logic with both epistemic modalities and justification assertions (cf. Artemov 2006; Artemov and Nogina 2005).

Corollary 6 (Model existence). *For any constant specification CS, J_{CS} is consistent and has a model.*

Proof. J_{CS} is consistent. Indeed, suppose J_{CS} proves $\bot$, and erase all justification terms (with ':'s) in each of its formulas. What remains is a chain of propositional formulas provable in classical logic (an easy induction on the length of the original proof) ending with $\bot$—contradiction.

To build a model for J_{CS}, use the Completeness Theorem (Theorem 4). Since J_{CS} does not prove $\bot$, by Completeness, there is a J_{CS}-model (where $\bot$ is false, of course). Q.E.D.

Factivity

Unlike Application and Monotonicity, *Factivity* of justifications is not required in basic Justification Logic systems, which makes the latter capable of representing both partial and factive justifications.

Factivity states that justifications of F are factive, i.e., sufficient for an agent to conclude that F is true. This yields the Factivity Axiom

$$t : F \to F, \tag{32.16}$$

which has a similar motivation to the Truth Axiom in epistemic modal logic

$$\mathbf{K}F \to F, \tag{32.17}$$

widely accepted as a basic property of knowledge (Plato, Wittgenstein, Hintikka, etc.).

The Factivity Axiom (32.16) first appeared in the Logic of Proofs LP as a principal feature of mathematical proofs. Indeed, in this setting (32.16) is valid: if there is a mathematical proof t of F, then F must be true.

We adopt the Factivity Axiom (32.16) for justifications that lead to knowledge. However, factivity alone does not warrant knowledge, which has been demonstrated by Russell and Gettier examples (Russell 1912; Gettier 1963)

Logic of Factive Justifications:

$$\mathsf{JT}_0 = \mathsf{J}_0 + \text{A4},$$

$$\mathsf{JT} = \mathsf{J} + \text{A4},$$

with

A4. *Factivity Axiom* $t : F \to F$.

Systems JT_{CS} corresponding to Constant Specifications *CS* are defined as in section "Logical Awareness and Constant Specifications".

JT-models are J-models with reflexive accessibility relations R. The reflexivity condition makes each possible world accessible from itself which exactly corresponds to the Factivity Axiom. The direct analogue of Theorem 3 holds for JT_{CS} as well.

Theorem 7. *For any Constant Specification CS, each of the logics* JT_{CS} *is sound and complete with respect to the class of* JT*-models respecting CS.*

Proof. We now proceed as in the proof of Theorem 4. The only addition to soundness is establishing that the Factivity Axiom holds in reflexive models. Let R be reflexive. Suppose $u \Vdash t : F$. Then $v \Vdash F$ for all v such that uRv. By reflexivity of R, uRu, hence $u \Vdash F$ as well.

For completeness, it suffices to check that R in the canonical model is reflexive. Indeed, if $s : F \in \Gamma$, then, by the properties of the maximal consistent sets, $F \in \Gamma$ as well, since JT derives $s : F \to F$ (with any *CS*). Hence $\Gamma^\sharp \subseteq \Gamma$ and $\Gamma R \Gamma$. Q.E.D.

Mkrtychev JT-models are singleton JT-models, i.e., JT-models with singleton W's.

Theorem 8. *For any Constant Specification CS, each of the logics* JT_{CS} *is sound and complete with respect to the class of Mkrtychev* JT*-models respecting CS.*

Proof. Soundness follows from Theorem 7. For completeness, we follow the footprints of Theorems 4 and 5, but define the accessibility relation R as

$$\Gamma R \Delta \ \textit{iff} \ \Gamma = \Delta.$$

Q.E.D.

Russell's Example: Induced Factivity

Here is Russell's well-known example from Russell (1912) of an epistemic scenario which can be meaningfully analyzed in Justification Logic.

> *If a man believes that the late Prime Minister's last name began with a 'B,' he believes what is true, since the late Prime Minister was Sir Henry Campbell Bannerman.*[4] *But if he believes that Mr. Balfour was the late Prime Minister, he will still believe that the late Prime Minister's last name began with a 'B,' yet this belief, though true, would not be thought to constitute knowledge.*

As in the Red Barn Example (section "Red Barn Example and Tracking Justifications"), we have to handle a wrong reason for a true justified fact. Again, the tools at Justification Logic seem to be useful and adequate here.

Let B stand for

the late Prime Minister's last name began with a 'B.'

Furthermore, let w be a wrong reason for B and r the right (hence factive) reason for B. Then, Russell's example yields the following assumptions:

$$\{w : B,\ r : B,\ r : B \to B\}. \tag{32.18}$$

In the original setting (32.18), we do not claim that w is a factive justification for B; moreover, such factivity is not completely consistent with our intuition. Paradoxically, however, in the basic Justification Logic J, we can logically deduce factivity of w from (32.18):

1. $r : B$—*an assumption*;
2. $r : B \to B$—*an assumption*;
3. B—*from 1 and 2, by Modus Ponens*;
4. $B \to (w : B \to B)$—*a propositional axiom*;
5. $w : B \to B$—*from 3 and 4, by Modus Ponens*.

[4] Which was common knowledge back in 1912.

However, this derivation utilizes the fact that r is a factive justification for B to conclude $w : B \to B$, which constitutes the case of 'induced factivity' of $w : B$. The question is, how can we distinguish the 'real' factivity of $r : B$ from an 'induced factivity' of $w : B$? Again, some sort of truth-tracking is needed here, and Justification Logic seems to do the job. The natural approach would be to consider the set of assumptions (32.18) **without** $r : B$, i.e.,

$$\{w : B,\ r : B \to B\}, \tag{32.19}$$

and establish that factivity of w, i.e., $w : B \to B$ is not derivable from (32.19). Here is a J-model $\mathcal{M} = (W, R, \mathcal{E}, \Vdash)$ in which (32.19) holds but $w : B \to B$ does not.

$W = \{\mathbf{0}\}$, $R = \emptyset$, $\mathbf{0} \nVdash B$, and $\mathcal{E}(t, F)$ holds for all pairs (t, F) except (r, B). It is easy to see that the closure conditions *Application* and *Sum* on $\mathcal{E}$ are fulfilled. At $\mathbf{0}$, $w : B$ holds, that is,

$$\mathbf{0} \Vdash w : B,$$

since w is an admissible evidence for B at $\mathbf{0}$ and there are no possible worlds accessible from $\mathbf{0}$. Furthermore,

$$\mathbf{0} \nVdash r : B,$$

since, according to $\mathcal{E}$, r is not an admissible evidence for B at $\mathbf{0}$. Hence

$$\mathbf{0} \Vdash r : B \to B.$$

On the other hand,

$$\mathbf{0} \nVdash w : B \to B$$

since B does not hold at $\mathbf{0}$.

Additional Principles and Systems

In this section, we discuss other principles and operations which may or may not be added to the core Justification Logic systems.

Positive Introspection

One of the common principles of knowledge is identifying *knowing* and *knowing that one knows*. In the formal modal setting, this corresponds to

$$\mathbf{K}F \to \mathbf{KK}F.$$

This principle has an adequate explicit counterpart: the fact that the agent accepts t as a sufficient evidence of F serves as a sufficient evidence that $t : F$. Often, such meta-evidence has a physical form, e.g., a referee report certifying that a proof of a paper is correct, a computer verification output given a formal proof t of F as an input, a formal proof that t is a proof of F, etc. *Positive Introspection* assumes that given t, the agent produces a justification $!t$ of $t : F$ such that

$$t : F \to !t : (t : F).$$

Positive Introspection in this operational form first appeared in the Logic of Proofs LP (Artemov 1995, 2001). A similar suggestion was made by Gödel (1995).

We define

$$\mathsf{J4} = \mathsf{J} + \mathrm{A5}$$

and

$$\mathsf{LP} = \mathsf{JT} + \mathrm{A5},^{5}$$

with

A5. *Positive Introspection Axiom* $t : F \to !t : (t : F)$.

We also define $\mathsf{J4}_0$, $\mathsf{J4}_{CS}$, LP_0, and LP_{CS} in the natural way (cf. section "Logical Awareness and Constant Specifications"). The direct analogue of Theorem 3 holds for $\mathsf{J4}_{CS}$ and LP_{CS} as well.

Note that in the presence of the Positive Introspection Axiom, one could limit the scope of the Axiom Internalization Rule R4 to internalizing axioms which are not yet of the form $e : A$. This is how it has been done in LP: the Axiom Internalization can then be emulated by using $!!e : (!e : (e : A))$ instead of $e_3 : (e_2 : (e_1 : A))$, etc. The notion of Constant Specification could also be simplified accordingly.

Such modifications are minor and they do not affect the main theorems and applications of Justification Logic.

Negative Introspection

Pacuit and Rubtsova considered in Pacuit (2005, 2006) and Rubtsova (2005, 2006) the *Negative Introspection* operation '?' which verifies that a given justification assertion is false. A possible motivation for considering such an operation could be that the positive introspection operation '!' may well be regarded as capable

[5] In our notation, LP can be assigned the name JT4. However, in virtue of a fundamental role played by LP for Justification Logic, we suggest keeping the name LP for this system.

of providing conclusive verification judgments about the validity of justification assertions $t : F$. So, when t is not a justification for F, such a '!' should conclude that $\neg t : F$. This is normally the case for computer proof verifiers, proof checkers in formal theories, etc. This motivation is, however, nuanced: the examples of proof verifiers and proof checkers work with both t and F as inputs, whereas the Pacuit-Rubtsova format $?t$ suggests that the only input for '?' is a justification t, and the result $?t$ is supposed to justify propositions $\neg t : F$ uniformly for all F's for which $t : F$ does not hold. Such an operation '?' does not exist for formal mathematical proofs since $?t$ should be a single proof of infinitely many propositions $\neg t : F$, which is impossible.[6] For what it's worth, we include Negative Introspection in the list of additional justification principles, and leave the decision of whether to accept it or not to the user.

A6. *Negative Introspection Axiom* $\neg t : F \to ?t : (\neg t : F)$.

We define systems

$$\mathsf{J45} = \mathsf{J4} + \text{A6},$$
$$\mathsf{JD45} = \mathsf{J45} + \neg t : \bot,$$
$$\mathsf{JT45} = \mathsf{J45} + \text{A4},$$

and naturally extend these definitions to $\mathsf{J45}_{CS}$, $\mathsf{JD45}_{CS}$, and $\mathsf{JT45}_{CS}$.

The direct analogue of Theorem 3 holds for $\mathsf{J45}_{CS}$, $\mathsf{JD45}_{CS}$, and $\mathsf{JT45}_{CS}$.

More Epistemic Models

We now define epistemic models for other Justification Logic systems.

- J4-models are J-models with *transitive* R and two additional conditions:
  Monotonicity with respect to R, i.e., $u \in \mathcal{E}(t, F)$ and uRv yield $v \in \mathcal{E}(t, F)$,
  Introspection closure: $\mathcal{E}(t, F) \subseteq \mathcal{E}(!t, t : F)$;
- LP-models are J4-models with *reflexive* R (these are the original Kripke-Fitting models);
- J45-models are J4-models satisfying conditions:
  Negative Introspection closure: $[\mathcal{E}(t, F)]^c \subseteq \mathcal{E}(?t, \neg t : F)$ (Here $[X]^c$ denotes the complement of X.)
  Strong Evidence: $u \Vdash t : F$ for all $u \in \mathcal{E}(t, F)$ (i.e., only 'actual' evidence is admissible).

[6] A proof-compliant way to represent negative introspection in Justification Logic was suggested in Artemov et al. (1999), but we will not consider it here.

Note that J45-models satisfy the *Stability* property: uRv yields '$u \in \mathcal{E}(t,F)$ iff $v \in \mathcal{E}(t,F)$.' In other words, $\mathcal{E}$ is monotone with respect to R^{-1} as well. Indeed, the direction '$u \in \mathcal{E}(t,F)$ yields $v \in \mathcal{E}(t,F)$' is due to Monotonicity. Suppose $u \notin \mathcal{E}(t,F)$. By Negative Introspection closure, $u \in \mathcal{E}(?t, \neg t : F)$. By Strong Evidence, $u \Vdash ?t : (\neg t : F)$. By the definition of forcing, $v \Vdash \neg t : F$, i.e., $v \nVdash t : F$. By Strong Evidence, $v \notin \mathcal{E}(t,F)$.

Note also that the Euclidean property of the accessibility relation R is not required for J45-models and is not needed to establish the soundness of J45 with respect to J45-models. However, the canonical model for J45 is Euclidean, hence both soundness and completeness claims trivially survive an additional requirement that R is Euclidean.

- JD45-models are J45-models with the *Serial* condition on the accessibility relation R: for each u there is v such that uRv holds.
- JT45-models are J45-models with reflexive R. Again, the Euclidean property (or, equivalently, symmetry) of R is not needed for soundness. However, these properties hold for the canonical JT45-model, hence they could be included into the formulation of the Completeness Theorem.

Theorem 9. *Each of the logics* $\mathsf{J4}_{CS}$, LP_{CS}, $\mathsf{J45}_{CS}$, $\mathsf{JT45}_{CS}$ *for any Constant Specification is sound and complete with respect to the corresponding class of epistemic models.* $\mathsf{JD45}_{CS}$ *is complete w.r.t. its epistemic models for axiomatically appropriate CS.*

Proof. We will follow the footprints of the proof of Theorem 4.

1. J4. For soundness, it now suffices to check the validity of the Positive Introspection Axiom at each node of any J4-model. Suppose $u \Vdash t : F$. Then $u \in \mathcal{E}(t,F)$ and $v \Vdash F$ for each v such that uRv. By the closure condition, $u \in \mathcal{E}(!t, t : F)$, and it remains to check that $v \Vdash t : F$. By monotonicity of $\mathcal{E}$, $v \in \mathcal{E}(t,F)$. Now, take any w such that vRw. By transitivity of R, uRw as well, hence $w \Vdash F$. Thus $v \Vdash t : F$, $u \Vdash\ !t :t :F$, and $u \Vdash t :F \rightarrow\ !t :t :F$.

   Completeness is again established as in Theorem 4. It only remains to check that the accessibility relation R is transitive, the admissible evidence function $\mathcal{E}$ is monotone, and the additional closure condition on $\mathcal{E}$ holds.

   Monotonicity. Suppose $\Gamma R \Delta$ and $\Gamma \in \mathcal{E}(t,F)$, i.e., $t : F \in \Gamma$. By maximality of Γ, $!t : t : F \in \Gamma$ as well, since $\mathsf{J4} \vdash t : F \rightarrow\ !t : t : F$. By definition, $t : F \in \Delta$, i.e., $\Delta \in \mathcal{E}(t,F)$.
   Transitivity. Suppose $\Gamma R \Delta$, $\Delta R \Sigma$, and $t : F \in \Gamma$. Then, by monotonicity, $t : F \in \Delta$. By the definition of R, $F \in \Sigma$, hence $\Gamma R \Sigma$.
   Closure. Suppose $\Gamma \in \mathcal{E}(t,F)$, i.e., $t : F \in \Gamma$. Then as above, $!t : t : F \in \Gamma$, hence $\Gamma \in \mathcal{E}(!t, t : F)$.

2. LP. This is the well-studied case of the Logic of Proofs, (cf. Fitting 2005).
3. J45. Soundness. We have to check the Negative Introspection Axiom. Let $u \Vdash \neg t : F$, i.e., $u \nVdash t : F$. By the Strong Evidence condition, $u \notin \mathcal{E}(t,F)$. By Negative Introspection closure, $u \in \mathcal{E}(?t, \neg t : F)$. By Strong Evidence, $u \Vdash ?t : (\neg t : F)$.

Completeness. We follow the same canonical model construction as in J and J4. The only addition is checking **Negative Introspection closure**. Let $\Gamma \notin \mathcal{E}(t,F)$. Then $t : F \notin \Gamma$. By maximality, $\neg t : F \in \Gamma$. By the Negative Introspection Axiom, $?t : (\neg t : F) \in \Gamma$, hence $\Gamma \in \mathcal{E}(?t, \neg t : F)$.

Here is an additional feature of the canonical model that can be included in the formulation of the Completeness Theorem to make it more specific.

R is Euclidean. Let $\Gamma R \Delta$ and $\Gamma R \Delta'$. It suffices to show that $\Delta^\sharp \subseteq \Delta'$. Let $F \in \Delta^\sharp$. Then for some t, $t : F \in \Delta$, i.e., $\Delta \in \mathcal{E}(t,F)$. By Stability, $\Gamma \in \mathcal{E}(t,F)$, hence $t : F \in \Gamma$ and $F \in \Gamma^\sharp$. By the definition of R, $F \in \Delta'$.

4. JD45. The proof can be found in Kuznets (2008).
5. JT45. For soundness, it suffices to check the Factivity Axiom, which easily follows from the reflexivity of R. For completeness, follow the footprints of 3 and note that R is reflexive. Indeed, $\Gamma^\sharp \subseteq \Gamma$ for reflexive theories.

   The additional features of the canonical model are as follows: *R is an equivalence relation, the admissible evidence function does not distinguish equivalent worlds*. This follows easily from 5.

Q.E.D.

Historical survey. The first Justification Logic system LP was introduced in 1995 in Artemov (1995) (cf. also Artemov 2001). Such basic properties of Justification Logic as internalization, realization, arithmetical semantics Artemov (1995, 2001), symbolic models and complexity estimates (Brezhnev and Kuznets 2006; Kuznets 2000; Milnikel 2007; Mkrtychev 1997), and epistemic semantics and completeness (Fitting 2003, 2005) were first established for LP.

A fair amount of work has already been done on Justification Logics other than LP. Systems J, J4, and JT were first considered in Brezhnev (2000) under different names and in a slightly different setting.[7] JT45 appeared independently in Pacuit (2005, 2006) and Rubtsova (2005, 2006), and JD45 in Pacuit (2005, 2006). J45 has, perhaps, first been considered in this work. Systems combining epistemic modalities and justifications were studied in Artemov (2006) and Artemov and Nogina (2004, 2005).

Mkrtychev semantics for J, JT, and J4 with Completeness Theorem were found in Kuznets (2000). Complexity bounds for LP and J4 were found in Kuznets (2000); Milnikel (2007). A comprehensive overview of all decidability and complexity results can be found in Kuznets (2008).

Forgetful Projection and the Correspondence Theorem

An intuitive connection between justification assertions and the justified belief modality $\Box$ involves the informal **existential quantifier**: $\Box F$ is read as

[7] Brezhnev (2000) also considered variants of Justification Logic systems which, in our notations, would be called "JD" and "JD4."

$$\textit{for some } x,\ x : F.$$

The language of Justification Logic does not have quantifiers over justifications, but instead has a sufficiently rich system of operations (polynomials) on justifications. We can use Skolem's idea of replacing quantifiers by functions and view Justification Logic systems as Skolemized logics of knowledge/belief. Naturally, to convert a Justification Logic sentence to the corresponding Epistemic Modal Logic sentence, one can use the **forgetful projection** '$\rightsquigarrow$' that replaces each occurrence of $t : F$ by $\Box F$.

Example: the sentence

$$x : P \to f(x) : Q$$

can be regarded as a Skolem-style version of

$$\exists x(x : P) \to \exists y(y : Q),$$

which can be read as

$$\Box P \to \Box Q,$$

which is the forgetful projection of the original sentence $x : P \to f(x) : Q$ (here, P, Q are assumed to be atomic sentences for simplicity's sake).

Examples (P, Q are atomic propositions):

$$t : P \to P \quad \rightsquigarrow \quad \Box P \to P,$$

$$t : P \to !t : (t : P) \quad \rightsquigarrow \quad \Box P \to \Box\Box P,$$

$$s : (P \to Q) \to (t : P \to (s \cdot t) : Q) \quad \rightsquigarrow \quad \Box(P \to Q) \to (\Box P \to \Box Q).$$

Forgetful projection sometimes forgets too much, e.g., a logical triviality $x : P \to x : P$, a meaningful principle $x : P \to (x + y) : P$, and a non-valid formula $x : P \to y : P$ have the same forgetful projection $\Box P \to \Box P$. However, '$\rightsquigarrow$' always maps valid formulas of Justification Logic to valid formulas of Epistemic Logic. The converse also holds: any valid formula of Epistemic Logic is a forgetful projection of some valid formula of Justification Logic. This follows from Correspondence Theorem 10. We assume that '$\rightsquigarrow$' is naturally extended from sentences to logics.

Theorem 10 (Consolidated Correspondence Theorem).

1. J $\rightsquigarrow$ K
2. JT $\rightsquigarrow$ T
3. J4 $\rightsquigarrow$ K4
4. LP $\rightsquigarrow$ S4
5. J45 $\rightsquigarrow$ K45
6. JD45 $\rightsquigarrow$ KD45
7. JT45 $\rightsquigarrow$ S5

Proof. It is straightforward that the forgetful projection of each of the Justification Logic systems J, JT, J4, LP, J45, JD45, JT45 is derivable in the corresponding epistemic modal logics K, T, K4, S4, K45, KD45, S5, respectively.

The core of Theorem 10 is the Realization Theorem:

> *One can recover justification terms for all modal operators in valid principles of epistemic modal logics* K, T, K4, S4, K45, KD45, *and* S5 *such that the resulting formula is derivable in the corresponding Justification Logic system* J, JT, J4, LP, J45, JD45, *and* JT45.

The important feature of the Realization Theorem is that it recovers realizing functions according to the **existential reading of the modality**, i.e., negative occurrences of the modality are realized by (distinct) free variables, and the positive occurrences by justification polynomials, depending on these variables. For example, $\Box F \to \Box G$ will be realized by $x : F' \to f(x) : G'$ where F', G' are realizations of F and G, respectively.

The Realization Theorem was first established for S4/LP (case 4) in Artemov (1995, 2001), cases 1–3 are covered in Brezhnev (2000). The Realization Theorem for 7 is established in Rubtsova (2006) using a very potent method from Fitting (2005), and the proof for 5 and 6 is very similar to Fitting (2005) and Rubtsova (2006) and can be safely omitted here. Q.E.D.

The Correspondence Theorem shows that the major epistemic modal logics K, K4, K45, KD45 (for belief) and T, S4, S5 (for knowledge) have exact Justification Logic counterparts J, J4, J45, JD45 (for partial justifications) and JT, LP, JT45 (for factive justifications).

Foundational Consequences of the Correspondence Theorem

Is there anything new that we have learned from the Correspondence Theorem about epistemic modal logics?

First of all, this theorem provides a new semantics for major modal logics. In addition to the traditional Kripke-style 'universal' reading of $\Box F$ as

F holds in all possible situations,

there is now a rigorous 'existential' semantics for $\Box F$ that reads as

there is a witness (proof, justification) for F.

Perhaps the justification semantics plays a similar role in modal logic to that played by Kleene realizability in intuitionistic logic. In both cases, the intended semantics was **existential**: the Brouwer-Heyting-Kolmogorov interpretation of intuitionistic logic (Heyting 1934; Troelstra and van Dalen 1988; van Dalen 1986) and Gödel's provability reading of S4 (Gödel 1933; Gödel 1995). In both cases, a later possible-world semantics of **universal** character became a highly potent and dominant technical tool. However, in both cases, Kripke semantics did not solve the

original semantical problems. It took Kleene realizability (Kleene 1945; Troelstra 1998) to reveal the computational semantics of intuitionistic logic and the Logic of Proofs (Artemov 1995, 2001) to provide exact BHK semantics of proofs for intuitionistic and modal logic.

In the epistemic context, Justification Logic and the Correspondence Theorem add a new 'justification' component to modal logics of knowledge and belief. Again, this new component was in fact an old and central notion which has been widely discussed by mainstream epistemologists but has remained out of the scope of formal logical methods. The Correspondence Theorem tells us that justifications are compatible with Hintikka-style systems and hence can be regarded as a foundation for epistemic modal logic.

Another comparison suggests itself here: Skolem functions for first-order logic which provide a functional reading of quantifiers. It might seem that Skolem functions do not add much, since they do not suggest altering first-order logic. However, Skolem functions proved to be very useful for foundations (e.g., Henkin and Herbrand models, etc.), as well as for applications (Resolution, Logic Programming, etc.).

Note that the Realization Theorem is not at all trivial. For cases 1–4, realization algorithms are known that use cut-free derivations in the corresponding modal logics (Artemov 1995, 2001; Brezhnev 2000; Brezhnev and Kuznets 2006). For 5–7, the Realization Theorem has been established by Fitting's method or its proper modifications (Fitting 2005; Rubtsova 2006). In principle, these results also produce realization procedures which are based on exhaustive search.

It would be a mistake to draw the conclusion that **any** modal logic has a reasonable Justification Logic counterpart. For example, the logic of formal provability GL (Artemov and Beklemishev 2005; Boolos 1993) contains the *Löb Principle*

$$\Box(\Box F \to F) \to \Box F, \tag{32.20}$$

which does not seem to have an epistemically acceptable explicit version. Let us consider, for example, a case when F is the propositional constant $\bot$ for *false*. A Skolem-style reading of (32.20) suggests that there are justification terms s and t such that

$$x : (s : \bot \to \bot) \to t : \bot. \tag{32.21}$$

This is intuitively false for factive justification, though. Indeed, $s : \bot \to \bot$ is the Factivity Axiom. Apply Axiom Internalization R4 to obtain $c : [s : \bot \to \bot]$ for some constant c. This choice of c makes the antecedent of (32.21) intuitively true and the conclusion of (32.21) false[8]. In particular, (32.20) is not valid for proof interpretation (cf. Goris 2007 for a total account of which principles of GL are realizable).

[8]To be precise, we have to substitute c for x everywhere in s and t.

Quantifier-Free First-Order Justification Logic

In this section, we extend J from the propositional language to the quantifier-free first-order language. To simplify formalities, we will regard here the first-order language without functional symbols, but with equality. Later, in section "Formalization of Gettier Examples", we will introduce definite descriptions in the form $\iota x F(x)$.

The language under consideration in this section is the first-order predicate language with individual variables and constants, predicate symbols of any arity and the equality symbol '=,' along with justification terms (including operations '$\cdot$' and '+') and the formula formation symbol ':' as in section "Basic Justification Logic J_0". Formulas are defined in the usual first-order way (without quantifiers) with an additional clause that if F is a formula and t is a justification polynomial, then $t : F$ is again a formula. The 'quantifier-free J' has all the axioms and rules of J, plus the equality axioms.

The formal system qfJ_0 has the following postulates:

A1. *Classical axioms of quantifier-free first-order logic with equality and Modus Ponens,*

A2. *Application Axiom* $s : (F \to G) \to (t : F \to (s \cdot t) : G)$,

A3. *Monotonicity Axiom* $s : F \to (s + t) : F$, $s : F \to (t + s) : F$,

E1. $g = g$ *for any individual term* g *(reflexivity of equality)*;

E2. $f = g \to (P[f/x] \to P[g/x])$ *(substitutivity of equality)*, where f and g are individual terms, P is any atomic formula, $P[f/x]$ and $P[g/x]$ are the results of replacing all the occurrences of a variable x in P by f and g respectively; we will use notations $P(f)$, $P(g)$ for that.

The system qfJ is $\mathsf{qfJ}_0 + \mathrm{R4}$, where

R4. *For each axiom* A *and any constants* $e_1, e_2, \ldots, e_n$, *infer* $e_n : e_{n-1} : \ldots : e_1 : A$.

As in section "Logical Awareness and Constant Specifications", we define Constant Specifications and systems qfJ_{CS}. In particular, $\mathsf{qfJ}_{\emptyset}$ is qfJ_0 and qfJ_{TCS} is qfJ.

The following proposition follows easily from the definitions.

Proposition 11. *Deduction Theorem holds for* qfJ_{CS} *for any constant specification CS. Internalization holds for* qfJ_{CS} *for an axiomatically appropriate constant specification CS.*

The following theorem provides a way to resolve the Frege puzzle Frege (1952) in an epistemic environment: equality of individual objects alone does not warrant substitutivity, but justified equality does.

Theorem 12 (Justified substitution). *For any individual terms* f *and* g, *justification variable* u, *and atomic formula* $P(x)$, *there is a justification term* $s(u)$ *such that* qfJ *proves*

$$u : (f = g) \to s(u) : [P(f) \leftrightarrow P(g)].$$

The same holds for any qfJ_{CS} *with an axiomatically appropriate constant specification CS.*

Proof. Taking into account Example 1, it suffices to establish that for some $t(u)$,

$$u : (f = g) \to t(u) : [P(f) \to P(g)].$$

From E2 it follows that qfJ proves

$$(f = g) \to [P(f) \to P(g)].$$

By R4, there is a justification constant c such that qfJ proves

$$c : \{(f = g) \to [P(f) \to P(g)]\}.$$

By A2, qfJ proves

$$c : \{(f = g) \to [P(f) \to P(g)]\} \to \{u : (f = g) \to (c \cdot u) : [P(f) \to P(g)]\}.$$

By Modus Ponens, qfJ proves

$$u : (f = g) \to (c \cdot u) : [P(f) \to P(g)].$$

It suffices now to pick $c \cdot u$ as $t(u)$. Q.E.D.

An unjustified substitution can fail in qfJ. Namely, for any individual variables x and y, a predicate symbol P, and justification term s, the formula

$$(x = y) \to s : [P(x) \leftrightarrow P(y)] \tag{32.22}$$

is not valid. To establish this, one needs some model theory for qfJ.

We define *qfJ-models* as the usual first-order Kripke models[9] equipped with admissible evidence functions. A model is $(W, \{D_w\}, R, \mathcal{E}, \Vdash)$ such that the following properties hold.

- W is an nonempty set of worlds.
- $\{D_w\}$ is the collection of nonempty domains D_w for each $w \in W$.
- R is the binary (accessibility) relation on W.
- $\mathcal{E}$ is the admissible evidence function which for each justification term t and formula F, returns the set of worlds $\mathcal{E}(t, F) \subseteq W$. Informally, these are the worlds

[9] Equality is interpreted as identity in the model.

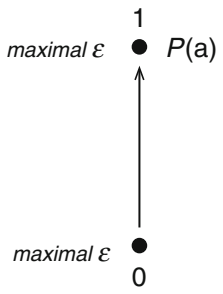


Fig. 32.1 Kripke-Fitting counter-model for unjustified substitution

where t is admissible evidence for F. We also assume that $\mathcal{E}$ satisfies the usual closure properties *Application* and *Sum* (section "Basic Epistemic Semantics").

- $\Vdash$ is the forcing (truth) relation such that

  $\Vdash$ assigns elements of D_w to individual variables and constants for each $w \in W$,
  for each n-ary predicate symbol P, and any $a_1, a_2, \ldots, a_n \in D_w$, it is specified whether $P(a_1, a_2, \ldots, a_n)$ holds in D_w,
  $\Vdash$ is extended to all the formulas by stipulating that
  $w \Vdash s = t$ iff '$\Vdash$' maps s and t to the same element of D_w,
  $w \Vdash P(t_1, t_2, \ldots, t_n)$ iff '$\Vdash$' maps t_i's to a_i's and $P(a_1, a_2, \ldots, a_n)$ holds in D_w,
  $w \Vdash F \wedge G$ iff $w \Vdash F$ and $w \Vdash G$,
  $w \Vdash \neg F$ iff $w \nVdash F$,
  $w \Vdash t : F$ iff $v \Vdash F$ for all v such that wRv, and $w \in \mathcal{E}(t, F)$.

The notion of a model respecting given constant specification is directly transferred from section "Basic Epistemic Semantics".

The following Theorem is established in the same manner as the soundness part of Theorem 4.

Theorem 13. *For any Constant Specification CS,* qfJ_{CS} *is sound with respect to the corresponding class of epistemic models.*

We are now ready to show that instances of unjustified substitution can fail in qfJ. To do this, it now suffices to build a qfJ-counter-model for (32.22) with the total constant specification. Obviously, the *maximal* $\mathcal{E}$ (i.e., $\mathcal{E}(t, F)$ contains each world for any t and F) respects any constant specification.

The Kripke-Fitting counter-model in Fig. 32.1 exploits the traditional modal approach to refute a belief assertion by presenting a possible world where the object of this belief does not hold. In the picture, only true atomic formulas are shown next to possible worlds.

- $W = \{\mathbf{0}, \mathbf{1}\}$; $R = \{(\mathbf{0}, \mathbf{1})\}$; $D_0 = D_1 = \{a, b\}$;
- $\mathbf{1} \Vdash P(a)$ and $\mathbf{1} \nVdash P(b)$; the truth value of P at $\mathbf{0}$ does not matter;
- x and y are interpreted as a at $\mathbf{0}$; x is interpreted as a and y as b at $\mathbf{1}$;
- $\mathcal{E}$ is maximal at $\mathbf{0}$ and $\mathbf{1}$.

Obviously, $\mathbf{0} \Vdash x = y$. Since $\mathbf{1} \nVdash P(x) \leftrightarrow P(y)$, for any justification term s, $\mathbf{0} \nVdash s : [P(x) \leftrightarrow P(y)]$. Hence

$$\mathbf{0} \nVdash x = y \rightarrow s : [P(x) \leftrightarrow P(y)].$$

Formalization of Gettier Examples

We consider Gettier's Case I in detail; Case II is much simpler logically and can be given similar treatment. We will present a complete formalization of Case I in qfJ with a definite description operation. Let

- $J(x)$ be the predicate *x gets the job*;
- $C(x)$ be the predicate *x has (ten) coins (in his pocket)*;
- `Jones` and `Smith` be individual constants denoting Jones and Smith, respectively[10];
- u be a justification variable.

Natural Model for Case I

Gettier's assumptions (d) and (e) contain a definite description

$$\textit{the man who will get the job.} \tag{32.23}$$

In this section, we will formalize Case I using a definite description ι-operation such that $\iota x P(x)$ is intended to denote

$$\textit{the } x \textit{ such that } P(x).$$

We interpret $\iota x P(x)$ in a given world of a qfJ-model as the element a such that $P(a)$ if there exists a unique a satisfying $P(a)$. Otherwise, $\iota x P(x)$ is undefined and any atomic formula where $\iota x P(x)$ actually occurs is taken to be false. Definite description terms are non-rigid designators: $\iota x P(x)$ may be given different interpretations in different worlds of the same qfJ-model (cf. Fitting 2007). The use of a definite description

$$\textit{Jones is the man who will get the job}$$

as a justified belief by Smith hints that Smith has strong evidence for the fact that at most one person will get the job. This is implicit in Gettier's assumption.

[10]Assuming that there are people seeking the job other than Jones and Smith does not change the analysis.

We now present a Fitting model $\mathcal{M}$ which may be regarded as an exact epistemic formulation of Case I.

1. At the actual world **0**, $J(\mathtt{Smith})$, $C(\mathtt{Smith})$, and $C(\mathtt{Jones})$[11] hold and $J(\mathtt{Jones})$ does not hold.
2. There is a possible belief world **1** for Smith at which $J(\mathtt{Jones})$ and $\neg J(\mathtt{Smith})$ hold. These conditions follow from proposition (d)

   Jones is the man who will get the job, and Jones has coins

   or, in logic form,

   $$(\mathtt{Jones} = \iota x J(x)) \ \wedge \ C(\mathtt{Jones})$$

   for which Smith has a strong evidence. In addition, Smith has no knowledge of 'Smith has coins' and there should be a possible world at which $C(\mathtt{Smith})$ is false; we use **1** to represent this possibility.
3. World **1** is accessible from **0**.
4. Smith has a *strong evidence of (d)*, which we will represent by introducing a justification variable u such that

   $$u : [(\mathtt{Jones} = \iota x J(x)) \wedge C(\mathtt{Jones})] \tag{32.24}$$

   holds at the actual world **0**. We further assume that the admissible evidence function $\mathcal{E}$ respects the justification assertion (32.24), which yields

   $$\mathbf{0} \in \mathcal{E}(u, (\mathtt{Jones} = \iota x J(x)) \wedge C(\mathtt{Jones})).$$

   To keep things simple, we can assume that $\mathcal{E}$ is the maximal admissible evidence function, i.e., $\mathcal{E}(t, F) = \{\mathbf{0}, \mathbf{1}\}$ for each t, F.

   These observations lead to the following model $\mathcal{M}$ on Fig. 32.2.

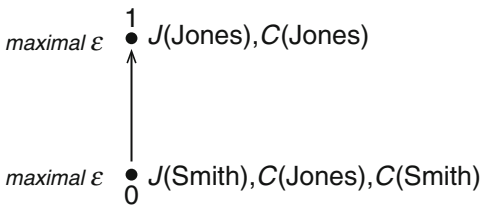


Fig. 32.2 Natural Kripke-Fitting model for Gettier Case I

[11] Strictly speaking, Case I explicitly states only that Smith has a strong evidence that $C(\mathtt{Jones})$, which is not sufficient to conclude that $C(\mathtt{Jones})$, since Smith's justifications are not necessarily factive. However, since the actual truth value of $C(\mathtt{Jones})$ does not matter in Case I, we assume that in this instance, Smith's belief that $C(\mathtt{Jones})$ was true.

- $W = \{\mathbf{0}, \mathbf{1}\}$; $R = \{(\mathbf{0}, \mathbf{1})\}$;
- $D_{\mathbf{0},\mathbf{1}} = \{\text{Jones}, \text{Smith}\}$, $\texttt{Jones}$ is interpreted as 'Jones' and $\texttt{Smith}$ as 'Smith';
- $\mathbf{0} \Vdash J(\texttt{Smith}),\ C(\texttt{Jones}),\ C(\texttt{Smith}), \neg J(\texttt{Jones})$;
- $\mathbf{1} \Vdash J(\texttt{Jones}),\ C(\texttt{Jones}),\ \neg J(\texttt{Smith}), \neg C(\texttt{Smith})$;
- $\iota x J(x)$ at $\mathbf{0}$ is interpreted as $\texttt{Smith}$ and at $\mathbf{1}$ as $\texttt{Jones}$;
- $\mathcal{E}$ is maximal at both $\mathbf{0}$ and $\mathbf{1}$.

It is interesting to compare this model with the axiomatic description of Case I. Here is the list of explicit assumptions:

$$J(\texttt{Smith}),\ C(\texttt{Smith}),\ C(\texttt{Jones}),\ \neg J(\texttt{Jones}),\ u : [(\texttt{Jones} = \iota x J(x)) \wedge C(\texttt{Jones})]. \tag{32.25}$$

It follows from the Soundness Theorem 13 that assumptions (32.25) provide a sound description of the actual world:

Proposition 14. $\mathsf{qfJ} + (32.25) \vdash F$ *entails* $\mathbf{0} \Vdash F$.

Example 15. The description of a model by (32.25) is not complete. For example, conditions (32.25) do not specifically indicate whether $t : C(\texttt{Smith})$ holds at the actual world for some t, whereas it is clear from the model that $\mathbf{0} \nVdash t : C(\texttt{Smith})$ for any t since $\mathbf{1} \nVdash C(\texttt{Smith})$ and $\mathbf{1}$ is accessible from $\mathbf{0}$. Model $\mathcal{M}$ extends the set of assumptions (32.25) to a possible **complete** specification: every ground proposition F in the language of this example is either true or false at the 'actual' world $\mathbf{0}$ of the model.

Formalizing Gettier's Reasoning

Gettier's conclusion in Case I states that *Smith is justified in believing that 'The man who will get the job has ten coins in his pocket.'* In our formal language, this amounts to a statement that for some justification term t,

$$t : C(\iota x J(x)) \tag{32.26}$$

is derivable in qfJ from assumptions of Case I.

Theorem 16. *Gettier's conclusion* $t : C(\iota x J(x))$ *is derivable in* qfJ *from assumptions (32.25) of Case I. Furthermore,* $t : C(\iota x J(x))$ *holds at the 'actual world'* $\mathbf{0}$ *of the natural model* $\mathcal{M}$ *of Case I.*

Proof. In order to find t we may mimic Gettier's informal reasoning. First, we formally derive (e) (i.e., $C(\iota x J(x))$) from (d) (i.e., $\texttt{Jones} = \iota x J(x) \wedge C(\texttt{Jones})$) and then use the fact that (d) is justified (i.e., $u : [\texttt{Jones} = \iota x J(x) \wedge C(\texttt{Jones})]$). We will now show that this argument can be formalized in qfJ. Note that in qfJ, we may reason as follows:

1. $\mathtt{Jones} = \iota x J(x) \rightarrow [C(\mathtt{Jones}) \rightarrow C(\iota x J(x))]$, *an axiom of* qfJ*;*
2. $[\mathtt{Jones} = \iota x J(x) \wedge C(\mathtt{Jones})] \rightarrow C(\iota x J(x))$, *by propositional reasoning, from 1;*
3. $s : \{[\mathtt{Jones} = \iota x J(x) \wedge C(\mathtt{Jones})] \rightarrow C(\iota x J(x))\}$, *by Internalization, from 2;*
4. $u : [\mathtt{Jones} = \iota x J(x) \wedge C(\mathtt{Jones})] \rightarrow (s \cdot u) : C(\iota x J(x))$, *by Axiom A2 and Modus Ponens, from 3;*
5. $u : [\mathtt{Jones} = \iota x J(x) \wedge C(\mathtt{Jones})]$, *an assumption from (32.25);*
6. $(s \cdot u) : C(\iota x J(x))$, *by Modus Ponens, from 4 and 5.*

Now we can pick t to be $s \cdot u$. So,

$$\mathsf{qfJ} + (32.25) \vdash (s \cdot u) : C(\iota x J(x))$$

and, by Proposition 14,

$$\mathbf{0} \Vdash (s \cdot u) : C(\iota x J(x)).$$

Q.E.D.

Eliminating Definite Descriptions, Russell-Style

We can eliminate definite descriptions from Case I using, e.g., Russell's translation (cf. Fitting and Mendelsohn 1998; Neale 1990; Russell 1905, 1919) of definite descriptions. According to Russell, $C(\iota x J(x))$ contains a hidden *uniqueness assumption* and reads as

$$\exists x [J(x) \wedge \forall y (J(y) \rightarrow y = x) \wedge C(x)], \tag{32.27}$$

and $\mathtt{Jones} = \iota x J(x)$ as

$$J(\mathtt{Jones}) \wedge \forall y (J(y) \rightarrow y = \mathtt{Jones}). \tag{32.28}$$

In addition, in the universe of Case I consisting of two objects *Jones, Smith*, a universally quantified sentence $\forall y F(y)$ reads as

$$F(\mathtt{Jones}) \wedge F(\mathtt{Smith}),$$

and an existentially quantified statement $\exists x G(x)$ reads as

$$G(\mathtt{Jones}) \vee G(\mathtt{Smith}).$$

Taking into account all of these simplifying observations, we may assume that for Smith (and the reader), $\forall y(J(y) \rightarrow y = \texttt{Jones})$ reads as

$$[J(\texttt{Jones}) \rightarrow (\texttt{Jones} = \texttt{Jones})] \wedge [J(\texttt{Smith}) \rightarrow (\texttt{Smith} = \texttt{Jones})],$$

which is equivalent[12] to

$$\neg J(\texttt{Smith}).$$

Now, (32.28) is equivalent to

$$J(\texttt{Jones}) \wedge \neg J(\texttt{Smith}),$$

and the whole Gettier proposition (d) collapses to

$$J(\texttt{Jones}) \wedge \neg J(\texttt{Smith}) \wedge C(\texttt{Jones}). \tag{32.29}$$

The assumption that (d) is justified for Smith can now be represented by

$$v : [J(\texttt{Jones}) \wedge \neg J(\texttt{Smith}) \wedge C(\texttt{Jones})], \tag{32.30}$$

for some justification variable v.

Smith's justified belief

'the man who will get the job has coins,' (32.31)

according to Russell, should read as

$$\exists x[J(x) \wedge \forall y(J(y) \rightarrow y = x) \wedge C(x)]. \tag{32.32}$$

The same considerations as above show that

$$\forall y[J(y) \rightarrow (y = \texttt{Jones})]$$

is equivalent to

$$\neg J(\texttt{Smith}),$$

and

$$\forall y[J(y) \rightarrow (y = \texttt{Smith})]$$

[12] We assume that everybody is aware that $\texttt{Smith} \neq \texttt{Jones}$.

is equivalent to

$$\neg J(\texttt{Jones}).$$

Since an existentially quantified formula $\exists x G(x)$ is logically equivalent to a disjunction $G(\texttt{Jones}) \vee G(\texttt{Smith})$, formula (32.32) is equivalent to

$$[J(\texttt{Jones}) \wedge \neg J(\texttt{Smith}) \wedge C(\texttt{Jones})] \vee [J(\texttt{Smith}) \wedge \neg J(\texttt{Jones}) \wedge C(\texttt{Smith})]. \tag{32.33}$$

Finally, the formalization of (32.31) in our language amounts to stating that for some justification term p,

$$p : \{[J(\texttt{Jones}) \wedge \neg J(\texttt{Smith}) \wedge C(\texttt{Jones})] \vee [J(\texttt{Smith}) \wedge \neg J(\texttt{Jones}) \wedge C(\texttt{Smith})]\}. \tag{32.34}$$

Theorem 17. *Gettier's claim (32.34) is derivable in* qfJ *from the assumption (32.30) of Case I, and holds in the 'actual world'* **0** *of the natural model* $\mathcal{M}$ *of Case I.*

Proof. After all the preliminary work and assumptions, there is not much left to do. We just note that (32.29) is a disjunct of (32.33). A derivation of (32.34) from (32.30) in qfJ reduces now to repeating steps of Example 2, which shows how to derive a justified disjunction from its justified disjunct. Q.E.D.

Comment 18. *One can see clearly the essence of Gettier's example. In (32.33), one of two disjuncts is justified but false, whereas the other disjunct is unjustified but true. The resulting disjunction (32.33) is both justified and true, but not really known to Smith.*

Hidden Uniqueness Assumption Is Necessary

In this subsection, we study what happens if we deviate from Russell's reading of definite descriptions, in particular if we skip the uniqueness of the defined object. For example, let us read Gettier's proposition (d) as

Jones will get the job, and Jones has ten coins in his pocket, (32.35)

and proposition (e) as

A man who will get the job has ten coins in his pocket. (32.36)

Then a fair formalization of (32.35) would be

$$J(\texttt{Jones}) \wedge C(\texttt{Jones}), \tag{32.37}$$

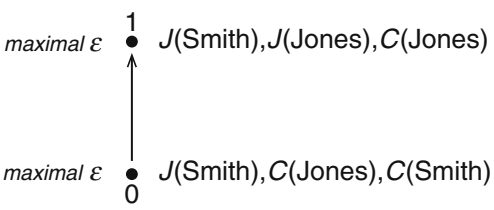


Fig. 32.3 Counter-model for Case I without uniqueness

and the assumption that (32.35) is justified for Smith is formalized as

$$u : [J(\texttt{Jones}) \wedge C(\texttt{Jones})]. \tag{32.38}$$

In this case, the set of explicitly made non-logical assumptions is

1. $u : [J(\texttt{Jones}) \wedge C(\texttt{Jones})]$, *assumption (32.38)*;
2. $\neg J(\texttt{Jones})$ *(Jones does not get the job)*;
3. $J(\texttt{Smith})$ *(Smith gets the job)*;
4. $C(\texttt{Smith})$ *(Smith has coins).*

Condition (32.36) naturally formalizes as

$$[J(\texttt{Jones}) \rightarrow C(\texttt{Jones})] \wedge [J(\texttt{Smith}) \rightarrow C(\texttt{Smith})]. \tag{32.39}$$

The claim that (32.39) is justified for Smith is formalized as

$$t : \{[J(\texttt{Jones}) \rightarrow C(\texttt{Jones})] \wedge [J(\texttt{Smith}) \rightarrow C(\texttt{Smith})]\} \tag{32.40}$$

for some justification term t.

We show that the assumptions 1–4 above do not suffice for proving (32.40).

Proposition 19. *For any justification term t, formula (32.40) is not derivable in* qfJ *from assumptions 1–4.*

Proof. Suppose (32.40) is derivable in qfJ from assumptions 1–4. Then, by the Deduction Theorem, qfJ would derive

$$\text{`}\textit{Conjunction of 1–4}\text{'} \rightarrow \textit{(32.40)}. \tag{32.41}$$

It now suffices to build a Fitting qfJ-model (Fig. 32.3) where (32.41) does not hold at a certain world.

At **0**, all assumptions 1–4 hold, but (32.40) is false at **0** for all t's. Indeed, (32.39) is false at **1**, since its conjunct

$$J(\texttt{Smith}) \rightarrow C(\texttt{Smith})$$

is false at **1**, and **1** is accessible from **0**. Q.E.D.

Streamlined Case I: No Coins/Pockets Are Needed

In this subsection, we show that references to coins and pockets, as well as definite descriptions, are redundant for making the point in Gettier example Case I. Here is a simpler, streamlined case based on the same material.

Smith has strong evidence for the proposition:

(d) Jones will get the job.
Proposition (d) entails:
(e) Either Jones or Smith will get the job.

Let us suppose that Smith sees the entailment from (d) to (e), and accepts (e) on the grounds of (d), for which he has strong evidence. In this case, Smith is clearly justified in believing that (e) is true. But imagine further that unknown to Smith, he himself, not Jones, will get the job. Then

(1) (e) is true,
(2) Smith believes that (e) is true, and
(3) Smith is justified in believing that (e) is true.

But it is equally clear that Smith does not know that (e) is true. . ..
In this version, the main assumption is

Smith has a strong evidence that Jones gets the job. (32.42)

Its straightforward formalization is

$$v : J(\texttt{Jones}). \tag{32.43}$$

The claim is that

Smith is justified in believing that either Jones or Smith will get the job. (32.44)

The natural formalization of the claim

$$t : [J(\texttt{Jones}) \vee J(\texttt{Smith})]. \tag{32.45}$$

The set of formal assumptions is

$$v : J(\texttt{Jones}),\ \ J(\texttt{Smith}),\ \ \neg J(\texttt{Jones}).$$

It is easy now to derive (32.45) in qfJ from assumption (32.43).

1. $v : J(\texttt{Jones})$, *assumption (32.43)*;
2. $J(\texttt{Jones}) \to J(\texttt{Jones}) \vee J(\texttt{Smith})$, *propositional axiom*;
3. $c : [J(\texttt{Jones}) \to J(\texttt{Jones}) \vee J(\texttt{Smith})]$, *from 2, by Axiom Internalization R4*;

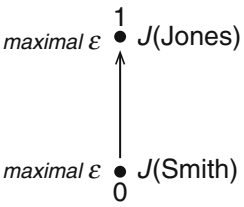


Fig. 32.4 Natural Kripke-Fitting model for the streamlined Case I

4. $c : [J(\texttt{Jones}) \to J(\texttt{Jones}) \vee J(\texttt{Smith})] \to [v : J(\texttt{Jones}) \to (c \cdot v) : (J(\texttt{Jones}) \vee J(\texttt{Smith}))]$, *Axiom A2*;
5. $(c \cdot v) : [J(\texttt{Jones}) \vee J(\texttt{Smith})]$, *from 4, 3, and 1, by Modus Ponens twice*.

At the actual world **0**, both hold:

$$J(\texttt{Jones}) \vee J(\texttt{Smith}) \text{ (meaning } \textit{(e) is true})$$

and

$$(c \cdot v) : [J(\texttt{Jones}) \vee J(\texttt{Smith})] \text{(meaning } \textit{(e) is justified}).$$

The desired Gettier-style point is made on the same material but without the unnecessary use of quantifiers, definite descriptions, coins, and pockets (Fig. 32.4).

It is fair to note, however, that Gettier example Case II in Gettier (1963) does not have these kinds of redundancies and is logically similar to the streamlined version of Case I presented above.

Gettier Example and Factivity

Theorem 20. *Gettier assumptions (32.25) in Case I are inconsistent in Justification Logic systems with factive justifications.*

Proof. Here is an obvious derivation of a contradiction in qfJT from (32.25):

$u : [(\texttt{Jones} = \iota x J(x)) \wedge C(\texttt{Jones})]$, *by (32.24)*;
$\texttt{Jones} = \iota x J(x)$, *by the Factivity Axiom and some propositional logic*;
$(\texttt{Jones} = \iota x J(x)) \to J(\texttt{Jones})$, *an assumed natural property of definite descriptions*;
$J(\texttt{Jones})$, *by Modus Ponens. This contradicts the condition* $\neg J(\texttt{Jones})$ *from (32.25)*.

Q.E.D.

The question is, what we have learned about Justification, Belief, Knowledge, and other epistemic matters?

Within the domain of formal epistemology, we now have a basic logic machinery to study justifications and their connections with Belief and Knowledge. Formalizing Gettier is a case study that demonstrates the method.

We show that Gettier reasoning was formally correct, with some hidden assumptions related to definite descriptions. Gettier examples belong to the area of Justification Logic dealing with partial justifications and are inconsistent within Justification Logic systems of factive justifications and knowledge. All this, perhaps, does not come as a surprise to epistemologists. However, these observations show that models provided by Justification Logic behave in a reasonable manner.

For epistemology, these developments are furthering the study of justification, e.g., the search for the 'fourth condition' of the JTB definition of knowledge. Justification Logic provides systematic examples of epistemological principles such as Application, Monotonicity, Logical Awareness, and their combinations, which look plausible, at least, within the propositional domain. Further discussion on these and other Justification Logic principles could be an interesting contribution to this area.

Conclusions

Justification Logic extends the logic of knowledge by the formal theory of justification. Justification Logic has roots in mainstream epistemology, mathematical logic, computer science, and artificial intelligence. It is capable of formalizing a significant portion of reasoning about justifications. In particular, we have seen how to formalize Kripke, Russell, and Gettier examples in Justification Logic. This formalization has been used for the resolution of paradoxes, verification, hidden assumption analysis, and eliminating redundancies.

Among other known applications of Justification Logic, so far there are

- intended provability semantics for Gödel's provability logic S4 with the Completeness Theorem (Artemov 1995, 2001);
- formalization of Brouwer-Heyting-Kolmogorov semantics for intuitionistic propositional logic with the Completeness Theorem (Artemov 1995, 2001);
- a general definition of the Logical Omniscience property, rigorous theorems that evidence assertions in Justification Logic are not logically omniscient (Artemov and Kuznets 2006). This provides a general framework for treating the problem of logical omniscience;
- an evidence-based approach to Common Knowledge (so-called Justified Common Knowledge) which provides a rigorous semantics to McCarthy's 'any fool knows' systems (Antonakos 2007; Artemov 2006; McCarthy et al. 1978). Justified Common Knowledge offers formal systems which are less restrictive than the usual epistemic logics with Common Knowledge (Artemov 2006).
- analysis of Knower and Knowability paradoxes (Dean and Kurokawa 2007, 2010).

It remains to be seen to what extent Justification Logic can be useful for analysis of empirical, perceptual, and *a priori* types of knowledge. From the perspective of Justification Logic, such knowledge may be considered as justified by constants (i.e., atomic justifications). Apparently, further discussion is needed here.

Acknowledgements The author is very grateful to Walter Dean, Mel Fitting, Vladimir Krupski, Roman Kuznets, Elena Nogina, Tudor Protopopescu, and Ruili Ye, whose advice helped with this paper. Many thanks to Karen Kletter for editing this text. Thanks to audiences at the CUNY Graduate Center, Bern University, the Collegium Logicum in Vienna, and the 2nd International Workshop on Analytic Proof Systems for comments on earlier versions of this paper. This work has been supported by NSF grant 0830450, CUNY Collaborative Incentive Research Grant CIRG1424, and PSC CUNY Research Grant PSCREG-39-721.

References

Antonakos, E. (2007). Justified and common knowledge: Limited conservativity. In S. Artemov & A. Nerode (Eds.), *Logical Foundations of Computer Science. International Symposium, LFCS 2007, Proceedings*, New York, June 2007 (Lecture notes in computer science, Vol. 4514, pp. 1–11). Springer.

Artemov, S. (1995). Operational modal logic. Technical report MSI 95-29, Cornell University.

Artemov, S. (1999). Understanding constructive semantics. In *Spinoza Lecture for European Association for Logic, Language and Information*, Utrecht, Aug 1999.

Artemov, S. (2001). Explicit provability and constructive semantics. *Bulletin of Symbolic Logic, 7*(1), 1–36.

Artemov, S. (2006). Justified common knowledge. *Theoretical Computer Science, 357*(1–3), 4–22.

Artemov, S. (2007). On two models of provability. In D. M. Gabbay, M. Zakharyaschev, & S. S. Goncharov (Eds.), *Mathematical problems from applied logic II* (pp. 1–52). New York: Springer.

Artemov, S. (2008). Symmetric logic of proofs. In A. Avron, N. Dershowitz, & A. Rabinovich (Eds.), *Pillars of computer science, essays dedicated to Boris (Boaz) Trakhtenbrot on the occasion of his 85th birthday* (Lecture notes in computer science, Vol. 4800, pp. 58–71). Berlin/Heidelberg: Springer.

Artemov, S., & Beklemishev, L. (2005). Provability logic. In D. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic* (2nd ed., Vol. 13, pp. 189–360). Dordrecht: Springer.

Artemov, S., Kazakov, E., & Shapiro, D. (1999). Epistemic logic with justifications. Technical report CFIS 99-12, Cornell University.

Artemov, S., & Kuznets, R. (2006). Logical omniscience via proof complexity. In *Computer Science Logic 2006*, Szeged (Lecture notes in computer science, Vol. 4207, pp. 135–149).

Artemov, S., & Nogina, E. (2004). Logic of knowledge with justifications from the provability perspective. Technical report TR-2004011, CUNY Ph.D. Program in Computer Science.

Artemov, S., & Nogina, E. (2005). Introducing justification into epistemic logic. *Journal of Logic and Computation, 15*(6), 1059–1073.

Artemov, S., & Strassen, T. (1993). Functionality in the basic logic of proofs. Technical report IAM 93-004, Department of Computer Science, University of Bern, Switzerland.

Boolos, G. (1993). *The logic of provability*. Cambridge: Cambridge University Press.

Brezhnev, V. (2000). On explicit counterparts of modal logics. Technical report CFIS 2000-05, Cornell University.

Brezhnev, V., & Kuznets, R. (2006). Making knowledge explicit: How hard it is. *Theoretical Computer Science, 357*(1–3), 23–34.

Dean, W., & Kurokawa, H. (2007). From the knowability paradox to the existence of proofs. Manuscript (submitted to *Synthese*).

Dean, W., & Kurokawa, H. (2010). The knower paradox and the quantified logic of proofs. In A. Hieke (Ed.), *Austrian Ludwig Wittgenstein society. Synthese, 176*(2), 177–225.

Dretske, F. (1971). Conclusive reasons. *Australasian Journal of Philosophy, 49*, 1–22.

Dretske, F. (2005). Is knowledge closed under known entailment? The case against closure. In M. Steup & E. Sosa (Eds.), *Contemporary Debates in Epistemology* (pp. 13–26). Malden: Blackwell.

Fagin, R., & Halpern, J. Y. (1985). Belief, awareness, and limited reasoning: Preliminary report. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, (pp. 491–501). Los Altos, CA: Morgan Kaufman.

Fagin, R., & Halpern, J. Y. (1988). Belief, awareness, and limited reasoning. *Artificial Intelligence, 34*(1), 39–76.

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. (1995). *Reasoning about knowledge*. Cambridge: MIT Press.

Fitting, M. (2003). A semantics for the logic of proofs. Technical report TR-2003012, CUNY Ph.D. Program in Computer Science.

Fitting, M. (2005). The logic of proofs, semantically. *Annals of Pure and Applied Logic, 132*(1), 1–25.

Fitting, M. (2007). Intensional logic. Stanford Encyclopedia of Philosophy (http://plato.stanford.edu), Feb 2007.

Fitting, M., & Mendelsohn, R. L. (1998). *First-order modal logic*. Dordrecht: Kluwer Academic.

Frege, G. (1952). On sense and reference. In P. Geach & M. Black (Eds.), *Translations of the philosophical writings of Gottlob Frege*. Oxford: Blackwell.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis, 23*, 121–123.

Gödel, K. (1933). Eine Interpretation des intuitionistischen Aussagenkalkuls. *Ergebnisse Math. Kolloq., 4*, 39–40. English translation in: S. Feferman et al., (Eds.) (1986). *Kurt Gödel collected works* (Vol. 1, pp. 301–303). Oxford: Oxford University Press/New York: Clarendon Press.

Gödel, K. (1995) Vortrag bei Zilsel/Lecture at Zilsel's (*1938a). In S. Feferman, J. W. Dawson Jr., W. Goldfarb, C. Parsons, & R. M. Solovay (Eds.), *Unpublished essays and lectures* (Kurt Gödel collected works, Vol. III, pp. 86–113). Oxford University Press.

Goldman, A. (1967). A causal theory of knowing. *The Journal of Philosophy, 64*, 335–372.

Goris, E. (2007). Explicit proofs in formal provability logic. In S. Artemov & A. Nerode (Eds.), *Logical Foundations of Computer Science. International Symposium, LFCS 2007, Proceedings*, New York, June 2007 (Lecture notes in computer science, Vol. 4514, pp. 241–253). Springer.

Hendricks, V. F. (2003). Active agents. *Journal of Logic, Language and Information, 12*(4), 469–495.

Hendricks, V. F. (2005). *Mainstream and formal epistemology*. New York: Cambridge University Press.

Heyting, A. (1934). *Mathematische Grundlagenforschung. Intuitionismus. Beweistheorie*. Berlin: Springer.

Hintikka, J. (1962). *Knowledge and belief*. Ithaca: Cornell University Press.

Hintikka, J. (1975). Impossible possible worlds vindicated. *Journal of Philosophical Logic, 4*, 475–484.

Kleene, S. (1945). On the interpretation of intuitionistic number theory. *The Journal of Symbolic Logic, 10*(4), 109–124.

Krupski, N. V. (2006). On the complexity of the reflected logic of proofs. *Theoretical Computer Science, 357*(1), 136–142.

Krupski, V. N. (2001). The single-conclusion proof logic and inference rules specification. *Annals of Pure and Applied Logic, 113*(1–3), 181–206.

Krupski, V. N. (2006). Referential logic of proofs. *Theoretical Computer Science, 357*(1), 143–166.

Kuznets, R. (2000). On the complexity of explicit modal logics. In *Computer Science Logic 2000* (Lecture notes in computer science, Vol. 1862, pp. 371–383). Berlin/Heidelberg: Springer.

Kuznets, R. (2008). *Complexity issues in justification logic*. PhD thesis, CUNY Graduate Center. http://kuznets.googlepages.com/PhD.pdf.

Lehrer, K., & Paxson, T. (1969). Knowledge: Undefeated justified true belief. *The Journal of Philosophy, 66*, 1–22.

Luper, S. (2005). The epistemic closure principle. Stanford Encyclopedia of Philosophy.

McCarthy, J., Sato, M., Hayashi, T., & Igarishi, S. (1978). On the model theory of knowledge. Technical report STAN-CS-78-667, Stanford University.

Meyer, J. -J. Ch., & van der Hoek, W. (1995). *Epistemic logic for AI and computer science*. Cambridge: Cambridge University Press.

Milnikel, R. (2007). Derivability in certain subsystems of the logic of proofs is Π_2^p-complete. *Annals of Pure and Applied Logic, 145*(3), 223–239.

Mkrtychev, A. (1997). Models for the logic of proofs. In S. Adian & A. Nerode (Eds.), *Logical Foundations of Computer Science '97, Yaroslavl'* (Lecture notes in computer science, Vol. 1234, pp. 266–275). Springer.

Moses, Y. (1988). Resource-bounded knowledge. In M. Vardi (Ed.), *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, Pacific Grove, March 7–9, 1988 (pp. 261–276). Morgan Kaufmann Publishers.

Neale, S. (1990). *Descriptions*. Cambridge: MIT.

Nozick, R. (1981). *Philosophical explanations*. Cambridge: Harvard University Press.

Pacuit, E. (2005). A note on some explicit modal logics. In *5th Panhellenic Logic Symposium*, Athens, July 2005.

Pacuit, E. (2006). A note on some explicit modal logics. Technical report PP-2006-29, University of Amsterdam. ILLC Publications.

Parikh, R. (1987). Knowledge and the problem of logical omniscience. In Z. Ras & M. Zemankova (Eds.), *ISMIS-87 International Symposium on Methodology for Intellectual Systems* (pp. 432–439). North-Holland.

Rubtsova, N. (2005). Evidence- for based knowledge S5. In *2005 Summer Meeting of the Association for Symbolic Logic, Logic Colloquium '05*, Athens, 28 July–3 August 2005. Abstract. Association for Symbolic Logic. (2006, June). *Bulletin of Symbolic Logic, 12*(2), 344–345. doi:10.2178/bsl/1146620064.

Rubtsova, N. (2006). Evidence reconstruction of epistemic modal logic S5. In *Computer Science – Theory and Applications. CSR 2006* (Lecture notes in computer science, Vol. 3967, pp. 313–321). Springer.

Russell, B. (1905). On denoting. *Mind, 14*, 479–493.

Russell, B. (1912). *The problems of philosophy*. London: Williams and Norgate/New York: Henry Holt and Company.

Russell, B. (1919). *Introduction to mathematical philosophy*. London: George Allen and Unwin.

Stalnaker, R. C. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy, 12*, 133–163.

Troelstra, A. S. (1998). Realizability. In S. Buss (Ed.), *Handbook of proof theory* (pp. 407–474). Amsterdam: Elsevier.

Troelstra, A. S., & Schwichtenberg, H. (1996). *Basic proof theory*. Amsterdam: Cambridge University Press.

Troelstra, A. S., & van Dalen, D. (1988). *Constructivism in mathematics* (Vols. 1 & 2). Amsterdam: North–Holland.

van Dalen, D. (1986). Intuitionistic logic. In D. Gabbay & F. Guenther (Eds.), *Handbook of philosophical logic* (Vol. 3, pp. 225–340). Dordrecht: Reidel.

von Wright, G. H. (1951). *An essay in modal logic*. Amsterdam: North-Holland.

Yavorskaya (Sidon), T. (2006). Multi-agent explicit knowledge. In D. Grigoriev, J. Harrison, & E. A. Hirsch (Eds.), *Computer Science – Theory and Applications. CSR 2006* (Lecture notes in computer science, Vol. 3967, pp. 369–380). Springer.

Chapter 33
Learning Theory and Epistemology

Kevin T. Kelly

Introduction

Learning is the acquisition of new knowledge and skills. It spans a range of processes from practice and rote memorization to the invention of entirely novel abilities and scientific theories that extend past experience. Learning is not restricted to humans: machines and animals can learn, social organizations can learn, and a genetic population can learn through natural selection. In that broad sense, learning is adaptive change, whether in behavior or in belief.

Learning can occur through the receipt of unexpected information, as when a detective learns where the suspect resides from an anonymous informant. But it can also be a process whose arrival at a correct result is in some sense guaranteed before the new knowledge is acquired. Such a learning process may be said to be *reliable* at the time it is adopted. *Formal Learning Theory* is an a priori, mathematical investigation of that strategic conception of reliability. It does not examine how people learn or whether people actually know, but rather, how reliable any system, human or otherwise, could possibly be. Thus, learning theory is related to traditional psychological and epistemological issues, but retains its own, distinct emphasis and character.

Reliability is a notoriously vague concept, suggesting a disposition to acquire new knowledge or skill over a broad range of relevantly possible environments. Learning theory deals with the vagueness not by insisting on a single, sharp "explication" of reliability, but by studying a range of possible explications, no one of which is insisted upon. That approach subtly shifts the focus from intractable

K.T. Kelly (✉)
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: kk3n@andrew.cmu.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_33

debates about what reliability *is* to the more objective task of determining which precise sens*es* of reliability are achievable in a given, precisely specified learning *problem*.

A learning problem specifies (1) what is to learned, (2) a range of relevantly possible environments in which the learner must succeed, (3) the kinds of inputs those environments provide to the learner, (4) what it means to learn over a range of relevantly possible environments, and (5) the sorts of learning strategies that will be entertained as solutions. A learning strategy *solves* a learning problem just in case it is admitted as a potential solution by the problem and succeeds in the specified sense over the relevant possibilities. A problem is *solvable* just in case some admissible strategy solves it.

Solvability is the basic question addressed by formal learning theory. To establish a positive solvability result, one must construct an admissible learning strategy and prove that the strategy succeeds in the relevant sense. A negative result requires a general proof that every allowable learning strategy fails. Thus, the positive results appear "methodological" whereas the negative results look "skeptical". Negative results and positive results lock together to form a whole that is more interesting than the sum of its parts. For example, a learning method may appear unimaginative and pedestrian until it is shown that no method could do better (i.e., no harder problem is solvable). And a notion of success may sound too weak until it is discovered that some natural problem is solvable in that sense but not in the more ambitious senses one would prefer.

There are so many different parameters in a learning problem that it is common to hold some of them fixed (e.g., the notion of success) and to allow others to vary (e.g., the set of relevantly possible environments). A partial specification of the problem parameters is called a learning *paradigm* and any problem agreeing with these specifications is an *instance* of the paradigm.

The notion of a paradigm raises more general questions. After several solvability and unsolvability results have been established in a paradigm, a pattern begins to emerge and one would like to know what it is about the combinatorial structure of the solvable problems that makes them solvable. A rigorous answer to that question is called a *characterization theorem*.

Many learning theoretic results concern the relative difficulty of two paradigms. Suppose we change a parameter (e.g., success) in one paradigm to produce another paradigm. There will usually remain an obvious correspondence between problems in the two paradigms (e.g., identical sets of serious possibilities). A *reduction* of paradigm P to another paradigm P' transforms a solution to a problem in P' into a solution to the corresponding problem in P. Then one may say that P is *no harder* than P'. Inter-reducible paradigms are *equivalent*. Equivalent paradigms may employ intuitively different standards of success, but the equivalence in difficulty shows that the quality of information provided by the diverse criteria is essentially the same. Paradigm equivalence results may therefore be viewed as epistemic analogues of the conservation principles of physics, closing the door on the temptation to get something (more reliability) for nothing by fiddling with the notion of success.

Learning in Epistemology

Epistemology begins with the irritating stimulus of unlearnability arguments. For example, Sextus Empiricus records the classical problem of inductive justification as follows:

> [Dogmatists] claim that the universal is established from the particulars by means of induction. If this is so, they will effect it by reviewing either all the particulars or some of them. But if they review only some, their induction will be unreliable, since it is possible that some of the particulars omitted in the induction may contradict the universal. If, on the other hand, their review is to include all the particulars, theirs will be an impossible task, because particulars are infinite and indefinite (Sextus 1985): 105.

That argument may be modelled in the following *input stream paradigm*. An *input stream* is just an infinite sequence e of natural numbers encoding discrete "observations". By stage n of inquiry the learner has seen observations $e(0)$, $e(1)$, ... $e(n-1)$. An *empirical proposition* is a proposition whose truth or falsity depends only on the input stream, and hence may be identified with a set of input streams. A learning strategy *decides* a given empirical proposition *with certainty* just in case in each relevantly possible input stream, it eventually halts and returns the truth value of the proposition.

Let the hypothesis to be assessed be "zeros will be observed forever", which corresponds to the empirical proposition whose only element is the everywhere zero input stream. Let every Boolean-valued input stream be a relevant alternative. To show that no possible learning strategy decides the hypothesis with certainty over these alternatives, construct a "demonic strategy" for presenting input in response to the successive outputs of an arbitrary learning strategy in such a way that the learner fails to halt with the right answer on the input stream presented. The demon presents the learner with the everywhere zero sequence until the learner halts and returns "true". If that never happens, the learner fails on the everywhere zero input stream. If the learner halts with "true", there is another relevantly possible input stream that agrees with the everywhere zero input stream up to the present and that presents only ones thereafter. The demon then proceeds to present this alternative input stream, on which the learner has already halted with the wrong answer. So whatever the learner's strategy does, it fails on some relevantly possible input stream and hence does not decide the hypothesis with certainty. That is the simplest example of a negative learning theoretic argument.

The argument actually shows something stronger. *Verification* with certainty requires, asymmetrically, that the learner's strategy halt with the output "true" if the hypothesis under assessment is true and that the strategy always say "false" otherwise, possibly without ever halting. The preceding argument shows that the "zeros forever" hypothesis is not verifiable with certainty.

Karl Popper's falsificationist epistemology was originally based on the observation that although universal hypotheses cannot be verified with certainty, they can be *refuted* with certainty, meaning that a method exists that halts with "false" if the hypothesis is false and that always says "true" otherwise. In the "zeros

forever" example, the refutation method simply returns "true" until a nonzero value is observed and then halts inquiry with "false".

When reliability demands verification with certainty, there is no tension between the static concept of conclusive justification and the dynamical concept of reliable success, since convergence to the truth occurs precisely when conclusive justification is received. Refutation with certainty severs that tie: the learner reliably stabilizes to the truth value of h, but when h is true there is no time at which that guess is certainly justified. The separation of reliablity from complete justification was hailed as a major epistemological innovation by the American Pragmatists.[1] In light of that innovation, one may either try to invent some notion of *partial* empirical justification (e.g., a theory of *confirmation*), or one may, like Popper, side entirely with reliability.[2] Learning theory has nothing to say about whether partial epistemic justification exists or what it might be. Insofar as such notions are entertained at all, they are assessed either as components of reliable learning strategies or as extraneous constraints on admissible strategies that may make reliability more difficult or even impossible to achieve. Methodological principles with the latter property are said to be *restrictive*.[3]

"Hypothetico-deductivism" is sometimes viewed as a theory of partial inductive support (Glymour 1980), but it can also been understood as a strategy for *reducing* scientific discovery to hypothesis assessment (Popper 1968; Kemeny 1953; Putnam 1963). Suppose that the relevant possibilities are covered by a countable family of hypotheses, each of which is refutable with certainty and informative enough to be interesting. A *discovery* method produces empirical hypotheses in response to its successive observations. A discovery method *identifies* these hypotheses *in the limit* just in case, on each relevantly possible input stream, the method eventually stabilizes to some true hypothesis in the family. Suppose that there is an assessment method that refutes each hypothesis with certainty. The corresponding hypothetico-deductive method is constructed as follows. It enumerates the hypotheses (by "boldness", "abduction", "plausibility", "simplicity", or the order by which they are produced by "creative intuition") and outputs the first hypothesis in the enumeration that is not rejected by the given refutation method. That reduction has occurred to just about everyone who has ever thought about inductive methodology. But things needn't be quite so easy. What if the hypotheses aren't even refutable with certainty? Could enumerating the right hypotheses occasion computational difficulties? Those

[1] "We may talk of the *empiricist* and the *absolutist* way of believing the truth. The absolutists in this matter say that we not only can attain to knowing truth, but we can know when we have attained to knowing it; while the empiricists think that although we may attain it, we cannot infallibly know when." (James 1948: 95–96).

[2] "Of course theories which we claim to be no more than conjectures or hypotheses need no justification (and least of all a justification by a nonexistent 'method of induction', of which nobody has ever given a sensible description)." (Popper 1982: 79).

[3] Cf. section "A foolish consistency" below.

are just the sorts of questions of principle that are amenable to learning theoretic analysis, as will be seen below.

Another example of learning theoretic thinking in the philosophy of science is Hans Reichenbach's "pragmatic vindication" of the "straight rule" of induction (Reichenbach 1938). Reichenbach endorsed Richard Von Mises' frequentist interpretation of probability. The relative frequency of an outcome in an input stream at position n is the number of occurrences of the outcome up to position n divided by n. The *probability* of an outcome in an input stream is the limit of the relative frequencies as n goes to infinity. Thus, a probabilistic statement determines an empirical proposition: the set of all input streams in which the outcome in question has the specified limiting relative frequency.

To discover limiting relative frequencies, Reichenbach recommended using the *straight rule*, whose guess at the probability of an outcome is the currently observed relative frequency of that outcome. It is immediate by definition that if the relevant possibilities include only input streams in which the limiting relative frequency of an event type is defined, then following the straight rule *gradually identifies* the true probability value, in the sense that on each relevantly possible input stream, for each nonzero distance from the probability, the conjectures of the rule eventually stay within that distance.

If the straight rule is altered to output an open interval of probabilities of fixed width centered on the observed relative frequency, then the modified method evidently identifies a true interval in the limit (given that a probability exists). That is the same property that hypothetico-deductive inquiry has over countable collections of refutable hypotheses.

So are probability intervals refutable with certainty? Evidently not, for each finite input sequence is consistent with each limiting relative frequency: simply extend the finite sequence with an infinite input sequence in which the probability claim is true. Is there any interesting sense in which open probability intervals can be reliably assessed? Say that a learner *decides* a hypothesis *in the limit* just in case in each relevantly possible environment, the learner eventually stabilizes to "true" if the hypothesis is true and to "false" if the hypothesis is false. According to that notion of success, the learner is guaranteed to end up with the correct truth value, even though no relevantly possible environment affords certain verification or refutation. But even assuming that some limiting relative frequency exists, open probability intervals are not decidable even in that weak, limiting sense (Kelly 1996).[4] A learner

[4] This footnote has been added in 2015. The results that follow are valid for the Reichenbach-Von Mises view that probability is limiting relative frequency. But that is not the most natural way to relate learning theoretic results to statistical inference. A better idea is to view chance as an unanalyzed probability distribution over outcomes. Learners receive sequences of random samples as inputs. Hypotheses are sets of possible chance distributions. Let $0 < \alpha \leq 1$. A learner α-verifies hypothesis H iff, in each possible chance distribution p, hypothesis H is true in p exactly when there exists a sample size at which the learner outputs "true" with chance $1 - \alpha$. Then open intervals of probability are α-verifiable, for $0 < \alpha \leq 1$. All of the learning theoretic criteria of success discussed below can be re-interpreted probabilistically in a similar way.

verifies a hypothesis *in the limit* just in case, on each relevantly possible input stream, she converges to "true" if the hypothesis is true and fails to converge to "true" otherwise. That even weaker notion of success is "one sided", for when the hypothesis is true, it is only guaranteed that "false" is produced infinitely often (possibly at ever longer intervals).[5] Analogously, *refutation in the limit* requires convergence to "false" when the hypothesis is false and anything but convergence to "false" otherwise. It turns out that open probability intervals are verifiable but not decidable in the limit given that some probability (limiting relative frequency) exists.[6]

Identification in the limit is possible even when the possible hypotheses are merely verifiable in the limit. Indeed, identification in the limit is in general reducible to limiting verification, but the requisite reduction is a bit more complicated than the familiar hypothetico-deductive construction. Suppose we have a countable family of hypotheses covering all the relevant possibilities and a limiting verifier for each of these hypotheses. Enumerate the hypotheses so that each hypothesis occurs infinitely often in the enumeration. At a given stage of inquiry, find the first remaining hypothesis whose limiting verifier currently returns "true". If there is no such, output the first hypothesis and go to the next stage of inquiry. If there is one, output it and delete all hypotheses occurring prior to it from the hypothesis enumeration. It is an exercise to check that this method identifies a true hypothesis in the limit. So although limiting verification is an unsatisfying sense of reliable assessment, it sufficees for limiting identification. If the hypotheses form a partition, limiting verifiability of each cell is also necessary for limiting identification (Kelly 1996). So limiting verification is more important than it might first have appeared.

Neyman and Pearson justified their theory of statistical testing in terms of the frequentist interpretation of probability:

> It may often be proved that if we behave according to such a rule, then in the long run we shall reject *h* when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject *h* sufficiently often when it is false (Neyman and Pearson 1933: 142).

The significance level of a test is a fixed upper bound on the limiting relative frequency of false rejection of the hypothesis under test over all possible input streams. A test is "useless" if the limiting frequency of mistaken acceptances exceeds one minus the significance, for then one could have done better at reducing the limiting relative frequency of error by ignoring the inputs and flipping a coin biased according to the significance level. "Useful" testability can be viewed as

[5] If there were any schedule governing the rate at which the outputs "false" spread apart through time, that schedule could be used to produce a method that decides the hypothesis in the limit: the new rule outputs "false" until the simulated rule produces more "true" outputs than the schedule allows for. Thus, the potential for ever rarer "false" outputs when the hypothesis is false is crucial to the extra lenience of this criterion.

[6] Conjecturing "true" while the observed frequency is in the interval and "false" otherwise does suffice unless we exclude possible input streams in which the limiting relative frequency approaches its limit from one side, for all but finitely many stages along the input stream. A reliable method is presented in Kelly (1996).

a learning paradigm over input streams. How does it relate to the "qualitative" paradigms just discussed? It turns out that the existence of a useful test for a hypothesis is equivalent to the hypothesis being either verifiable or refutable in the limit (Kelly 1996). That is an example of a paradigm equivalence theorem, showing that useful statistical tests provide essentially no more "information" than limiting verification or refutation procedures, assuming the frequentist interpretation of probability.

It is standard to assume in statistical studies that the relevant probabilities exist, but is there a sense in which that claim could be reliably assessed? Demonic arguments reveal the existence of a limiting relative frequency to be neither verifiable in the limit nor refutable in the limit over arbitrary input streams. But that hypothesis is *gradually verifiable* in the sense that there is a method that outputs numbers in the unit interval such that the numbers approach one just if the hypothesis is true (Kelly 1996). A demonic argument shows that the existence of a limiting relative frequency is not *gradually refutable,* in the sense of producing a sequence of numbers approaching zero just in case that hypothesis is false.

Gradual decidability requires that the learner's outputs gradually converge to the truth value of the hypothesis whatever the truth value happens to be. Unlike gradual verification and refutation, which we have just seen to be weaker than their limiting analogues, gradual decision is inter-reducible with limiting decision: simply choose a cutoff value (e.g. 0.5) and output "true" if the current output is less than 0.5 and "false" otherwise. Gradual decision is familiar as the sense of success invoked in Bayesian convergence arguments. Since Bayesian updating by conditionalization can never retract a zero or a one on input of nonzero probability, those outputs indicate certainty (inquiry may as well be halted), so limiting decision may only be accomplished gradually.

This short discussion illustrates how familiar epistemological issues as diverse as the problem of induction, Popper's falsificationism, Reichenbach's vindication of the straight rule, statistical testability, and Bayesian convergence all fit within a single, graduated system of learnability concepts.

Computable Learning

The preceding discussion frames traditional epistemological topics in learning theoretic terms. But unlike traditional epistemology, formal learning theory develops systematic connections between inductive inference and computation.

One of the earliest examples of a computationally driven unlearnability argument was presented by Hilary Putnam in 1963 in an article criticizing Rudolph Carnap's (1950) approach to inductive logic. Following suggestions by Wittgenstein, Carnap viewed inductive logic as a theory of "partial entailment", in which the conditional probability of the hypothesis given the inputs is interpreted as the proportion of logical possibilities satisfying the "premise" that also satisfy the intended "conclusion".

An inductive logic determines a *prediction* function: given the inputs encountered so far, output the most probable guess at the next datum to be seen. Interpret a tie as a refusal to select a prediction and count it as a failure at that round. Since the relevant probabilities are computable in Carnap's inductive logic, so is the induced prediction function.

In the *extrapolation paradigm*, the goal in each relevantly possible input stream is to eventually produce only correct predictions. Putnam showed that no computable prediction function can extrapolate the set of all total computable input streams, from which it follows that Carnap's inductive logic cannot extrapolate the computable input streams. Let an arbitrary, computable prediction strategy be given. At each stage, the demon calculates the computable prediction strategy's next prediction in light of the input already presented. If the prediction is one or greater, the demon presens a zero. If the prediction is zero, the demon presents a one. Evidently, every prediction made by the computable extrapolator along the resulting input stream is wrong. Since both the demon's strategy and the learner's strategy are both computable. The input stream presented by the demon is computable and, hence, relevantly possible.[7]

On the other hand, the problem is solved by an obvious, but uncomputable, hypothetico-deductive method. Enumerate a set of computer programs that compute all and only the total computable functions (i.e., no programs that go into infinite loops are included). Each such program is computably refutable with certainty, by calculating its prediction for the current stage of inquiry and rejecting it if the prediction does not agree with what is observed. That method identifies a correct program in the limit. To turn it into a reliable extrapolator, just compute what the currently output hypothesis says will happen at the next stage (another example of a paradigm reduction).

The only part of that procedure that is not computable is enumerating a collection of programs covering exactly the total computable functions. Since the prediction problem is computably unsolvable, it follows immediately that no such program enumeration is computable. So computable predictors fail on this problem "because" they cannot enumerate the right collection of hypotheses.[8]

The *computable function identification* paradigm poses the closely related problem of identifying in the limit a computer program correctly predicting each position in the input stream. The preceding hypothetico-deductive method uncomputably identifies the computable input streams in that sense, but in a seminal paper, E. M. Gold (1965) showed that the problem is not computably solvable. The computable demonic construction employed in the proof of Gold's result is more subtle than in

[7] Putnam's actual argument was more complicated than the version presented here.

[8] Putnam concluded that a scientific method should always be equipped with an extra input slot into which hypotheses that occur to us during the course of inquiry can be inserted. But such an "open minded" method must hope that the external hypothesis source (e.g., "creative intuition") does not suggest any programs that go into infinite loops, since the inability to distinguish such programs from "good" ones is what restricted the reliability of computable predictors to begin with!

the extrapolation case, because it is a nontrivial matter for a computable demon to figure out what the computable learner's current hypothesis predicts the next datum to be. For all the computable demon knows, the prediction may be undefined (i.e., the hypothesis may go into an infinite loop).

The demon proceeds in stages as follows[9]: At stage n, the learner has received a sequence of n inputs. The demon employs a fixed, computable enumeration of all ordered pairs of natural numbers. He then seeks the first pair (i, j) in the enumeration such that, after reading the current input followed by i zeros, the learner outputs a program that halts in j steps of computation with a prediction of zero for the next datum. If the search terminates with some such pair (i, j), then the demon adds i zeros to the input presented so far, and then presents a one (falsifying the hypothesis output by the learner after seeing the last zero). Otherwise, the demon continues searching forever and never proceeds to the next stage.

Suppose that the demon's construction runs through infinitely many stages. Then the search for a pair always terminates, so the resulting input stream falsifies the learner's conjecture infinitely often. The input stream is computable because it is produced by the interaction of two computable strategies. Suppose, then, that the demon's construction eventually gets stuck at a given stage. Then the demon's search for a pair fails. So on the input stream consisting of the input presented so far followed by all zeros, the learner never produces a hypothesis that correctly predicts the next zero. Thats input stream is also computable: use a finite lookup table to handle the input presented so far and output zero thereafter. So in either case, the demon never identifies a correct program along some relevantly possible input stream.

Since the demon makes the learner's conjecture false infinitely often, his strategy wins even if we weaken the criterion of success to *unstable* identification in the limit, according to which the learner must eventually output only true hypotheses, but need not stabilize to a particular hypothesis.[10]

Each total computer program is computably refutable with certainty (compute its successive predictions and compare them to the input), so we now know that computable refutability with certainty reduces neither computable extrapolation nor computable limiting identification. Does computable identification in the limit reduce computable extrapolation? One might suppose so: just compute the prediction of the limiting identifier's current conjecture, which must eventually be right since the identifier's conjectures are eventually correct. But although the limiting identifier eventually produces programs without infinite loops, nothing prevents it from producing defective programs in the short run. If a computer attempts to derive

[9]This construction (Case and Smith 1983) is a bit stronger than Gold's. It produces an input stream on which infinitely many outputs of the learner are wrong. Gold's construction merely forces the learner to vacillate forever (possibly among correct conjectures).

[10]Cf. the preceding footnote. In the learning theoretic literature, unstable identification is called BC identification for "behaviorally correct", whereas stable identification is called EX identification for "explanatory". Osherson and Weinstein (1986) call stable identification "intensional" and unstable identification "extensional".

predictions from those conjectures in the manner just described, it may get caught in an infinite loop and hang for eternity.

Blum and Blum (1975) constructed a learning problem that is computably identifiable in the limit but not computably extrapolable for just that reason. Consider a problem in which an unknown Turing machine without infinite loops is hidden in a box and the successive input are the (finite) runtimes of that program on successive inputs. The learner's job is to guess some computer program whose runtimes match the observed runtimes for each input (a task suggestive of fitting a computational model to psychological reaction time input). In that problem, every program is computably refutable with certainty: simulate it and see if it halts precisely when the input say it should. Infinite loops are no problem, for one will observe in finite time that the program doesn't halt when it should have. Since the set of all programs is computably enumerable (we needn't restrict the enumeration to *total* programs this time), a computable implementation of the hypothetico-deductive strategy identifies a correct hypothesis in the limit. Nonetheless, computable extrapolation of runtimes is not possible. Let a computable extrapolator be given. The demon is a procedure that wastes computational cycles in response to the computable predictor's last prediction. So at a given stage, the demonic program simulates the learner's program on the successive runtimes of the demonic program on earlier inputs. Whatever the learner's prediction is, the demon goes into a wasteful subroutine that uses at least one more step of computation than the predictor expected.

Another question raised by the preceding discussion is whether stable identification is equivalent to or harder than unstable identification for computable learners in the computable function identification paradigm. That question is answered affirmatively by Case and Smith (1983). To see why the answer might be positive, consider the function identification problem in which the relevant possibilities are the "almost self-describing input streams". A unit variant of an input stream is a partial computable function that is just like the input stream except that it may disagree or be undefined in at most one position. An input stream is *almost self-describing* just in case it is a unit variant of the function computed by the program whose index (according to a fixed, effective encoding of Turing programs into natural numbers) occurs in the input stream's first position. In other words, an "almost self-describing" input stream "gives away" a nearly correct hypothesis, but it doesn't say where the possible mismatch might be. An unstable learner can succeed by continually patching the "given away" program with ever larger lookup tables specifying what has been seen so far, since eventually the lookup table corrects the mistake in the "given away" program. But a stable learner would have to know *when* to stop patching, and that information was not given away.

In the problem just described, it is trivial to stably identify an almost correct program (just output the first datum) whereas no computable learner can stably identify an exactly correct program. Indeed, for each finite number of allowed errors there is a learning problem that is computably solvable under that error error fewer but not with one error fewer (Case and Smith 83). That result, known as the *anomaly hierarchy theorem,* can be established by means of functions that are self-describing up to n possible errors.

There are many more sophisticated results of the kind just presented, all of which share the following points in common. (1) Uncomputability is taken just as seriously as the problem of induction from the very outset of the analysis. That is different from the approach of traditional epistemology, in which idealized logics of justification are proposed and passed along to experts in computation for advice on how to satisfy them (e.g., Levi 1991). (2) When computability is taken seriously, the *halting problem* (the *formal* problem of determining whether a computer program is in an infinite loop on a given input) is very similar to the classical problem of induction: for as soon as one is sure that a computation will never end, it might, for all the simulator knows *a priori,* halt at the next stage. (3) Thus, computable learners fail when ideal ones succeed because computable solvability requires the learner to solve an *internalized* problem of induction (Kelly and Schulte 1997).

Some Other Paradigms

E. M. Gold's *language learnability* paradigm (1967) was intended to model child language acquition. In that setting, a *language* is just a computably enumerable set and a hypothesis is a code number (index) of a procedure that *accepts* all and only the members of the set.[11] Different kinds of relevantly possible environments are considered. An *informant* for a language is an enumeration of all possible strings labelled as positive or negative examples of the language. A *text* for a language is an enumeration of the elements of the language, and hence provides only positive information about membership.

Gold showed a number of results that attracted attention in cognitive science. The results for informants are similar to those for computable function identification. For example, (1) the obvious hypothetico-deductive method (non-computably) identifies all languages and (2) even the set of all computably decidable languages is not computably identifiable in the limit (the proof is similar to the one showing that the total computable functions are not identifiable in the limit). But the results for text are much weaker. For example, no collection of languages containing one infinite language and all finite subsets of that language is identifiable in the limit, even by non-computable learners.[12] Since children seem to learn language with few negative examples or corrections (Brown and Hanlon 1970), there have been attempts to obtain stronger positive results. For example, Wexler and Culicover (1980) modelled the environment as a presentation of context-utterance pairs, exchanging language learning from positive examples for the easier problem of computable function

[11] I.e., the procedure halts on members of the set (indicating acceptance) and not on any other inputs.

[12] The demon presents a text for the infinite language until the learner outputs a grammar for it, then keeps repeating the preceding datum until the learner produces a grammar for the input presented so far, then starts presenting the text from where he left off last, etc.

identification. Many other variations of the language learnability paradigm have been examined.[13]

The special difficulty with learning from text is "over-generalization", or leaping to a language that properly extends the actual language, for then no further input will correct the error. If there is no way to avoid positioning a language prior to one of its proper subsets (e.g., an infinite language must occur prior to all but finitely many of its finite subsets), hypothetico-deductivism must fail, since it will converge to the large language when one of its subsets is true. What is required is a way to use evidence to avoid overgeneralizing. That can be accomplished if (†) each possible language has a finite, characteristic sample such that once that sample is seen, the language can be produced without risk of overgeneralization. Then one may proceed by enumerating the relevantly possible grammars and conjecturing the first in the enumeration that is consistent with the input and whose characteristic sample has been observed. If no such grammar exists, stick with the preceding conjecture. Condition (†) is both necessary and sufficient for a collection of languages to be identifiable in the limit from text (Angluin 1980; Osherson and Weinstein 1986), providing our first example of a learning theoretic *characterization theorem*. Computable identification from text is characterized by the existence of a procedure that enumerates the characteristic sample for a language when provided with the index of a formal verification program for that language.

The *logical paradigm* (Shapiro 1981; Osherson and Weinstein 1986, 1989a, b; Kelly and Glymour 1989, 1990), situates learning theoretic ideas in a more traditional epistemological setting. In that paradigm, there is a first-order language in which to frame hypotheses and the underlying world is a countable relational structure interpreting that language. An environment consists of such a structure together with a variable assignment onto the domain of the structure and an enumeration of the set of all quantifier-free formulas true under that assignment.[14] The relevant possibilities are all the environments presenting models of some theory representing the learner's background knowledge.

An hypothesis assessment method tries to guess the truth value of a particular sentence or theory in light of the increasing information provided by the environment, and successful assessment can be interpreted in any of the senses introduced above. So for example, the dense order postulate (each pair of points has a point between them) is refutable but not verifiable in the limit given as background the theory of total orders with endpoints (Osherson and Weinstein 1989a, b).

The characterization theorem for this paradigm explains the grain of truth in the positivist's program of linking "cognitive significance" to logical form. An hypothesis is refutable (respectively, verifiable) with certainty given background theory K just in case the hypothesis is equivalent in K to a sentence in prenex

[13] A systematic compendium of results on language learnability is Osherson and Weinstein (1986).

[14] The "onto" assumption can be dropped if empirical adequacy rather than truth is the goal Lauth (1993).

normal form[15] with a purely universal (respectively, existential) quantifier prefix. Similarly, an hypothesis is refutable (respectively, verifiable) in the limit given K just in case it is equivalent in K to a prenex sentence with a prefix of form $\forall\exists$ (respectively, $\exists\forall$) (Osherson and Weinstein 1989a, b; Kelly and Glymour 1990). As one might expect, decision with certainty is possible just in case the hypothesis is equivalent to a quantifier-free sentence in K and decision in the limit (and hence gradual decision) is possible just in case the hypothesis is equivalent in K to a finite Boolean combination of purely universal and existential sentences.

A discovery method outputs theories in response to the information provided. As the goal of discovery, one can require that the method converge to the complete true theory in some fragment of the language (e.g., the purely universal sentences). *Uniform* theory identification requires that after some time the outputs of the method are true and entail the complete theory of the required fragment. For example, the complete truth is uniformly identifiable in the limit in a language with only unary predicates, but if there is a binary predicate or a unary predicate and a function symbol in the language, then neither the purely universal nor the purely existential fragment of the complete truth is identifiable in the limit (Kelly and Glymour 1989; Kelly 1996). *Nonuniform* or *pointwise* theory identification requires only that each true sentence in the specified fragment is eventually always entailed by the scientist's successive conjectures and that each false sentence is eventually never entailed. The theory of all true Boolean combinations of universal and existential sentences is identifiable in the limit in that sense. Thus, nonuniform theory identification provides a logical conception of scientific progress that, unlike Popper's "deductivist" epistemology, treats verifiable and refutable hypotheses symmetrically.

Nonuniform theory identification bears on another Popperian difficulty. Popper held that hypothetico-deductivism leads us ever closer to the truth in the limit. David Miller (1974) argued that "closeness" to the truth is not a semantic notion since it is not preserved under translation. Thomas Mormann (1988) traced the difficulty to mathematics: translation is a type of topological equivalence, but topological equivalence permits "stretching" and hence does not preserve distance (e.g., verisimilitude). Nonuniform identification is a topological rather than a metrical notion, and hence is preserved under translation, thereby avoiding Miller-style objections. Nonetheless it constitutes a nontrivial account of scientific progress toward the complete truth that does not imply that any future theory produced by science will be literally true.

[15] I.e., the sentence has the form of a quantifier-free sentence preceded by a sequence of quantifiers.

Reliability and Complexity

Learnability is a matter of how the possible futures making different hypotheses correct branch off from one another through time. The more complex the temporal entanglement of the futures satisfying incompatible hypotheses, the more difficult learning will be. Learnability is governed by the *topological* complexity of the possible hypotheses, and computable learnability depends on their *computational* complexity.[16]

Input streams can be topologized in an epistemologically relevant manner as follows. A *fan* of input streams is the set of all input streams extending some finite input sequence, which we may call the *handle* of the fan. A fan with a given handle is just the empirical proposition asserting that the handle has occurred in the input. An empirical proposition is *open* just in case it is a union of fans and is *closed* just in case its complement is open.[17] Then we have the following characterization: an empirical proposition is verifiable with certainty just in case it is open, is refutable with certainty just in case it is closed, and is decidable with certainty just in case it is both closed and open. For suppose that a hypothesis is open. To verify it with certainty, just wait until the observed input sequence is the handle of a fan contained in the hypothesis and halt inquiry with "true". Conversely, if a given method verifies a hypothesis with certainty, the hypothesis can be expressed as the union of all fans whose handles are finite input sequences on which the method halts with "true".

To characterize limiting and gradual success, topological generalizations of the open and closed propositions are required. Call the open and closed propositions the Σ_1 and Π_1 propositions, respectively. For each n, the Σ_{n+1} propositions are countable unions (disjunctions) of Π_n propositions and the Π_{n+1} propositions are countable intersections of Σ_n propositions. At each level n, a proposition is Δ_n just in case it is both Π_n and Σ_n. These are known as the *finite Borel* complexity classes, which have been familiar in functional analysis since early in this century (Hinman 1978). Then it can be shown that limiting verifiability, refutability, and decidability are characterized by Σ_2, Π_2, and Δ_2, respectively and that gradual verifiability, refutability, and decidability are characterized by Π_3, Σ_3 and $\triangle_2$, respectively. It can also be shown that when the hypotheses are mutually incompatible, stable identification in the limit is characterized by each hypothesis being Σ_2.[18]

In computable inquiry, attaching hypotheses to propositions is a nontrivial matter, so instead of bounding the complexity of empirical propositions, one must

[16]The computational versions of these ideas are in Gold (1965), Putnam (1965), and Kugel (1977). The topological space is introduced in Osherson and Weinstein (1986) and the characterizations are developed in Kelly (1992, 1996) Logical versions of the characterizations are developed in Osherson and Weinstein (1991) and Kelly and Glymour (1990).

[17]These are, in fact, the open sets of an extensively studied topological space known as the *Baire space* (Hinman 1978).

[18]Necessity of the condition fails if the hypotheses are mutually compatible or if we drop the stability requirement.

consider the overall *correctness relation* $C(e, h)$, which expresses that hypothesis h is correct in environment e. In computable function identification, for example, correctness requires that h be the index of a computer program that computes e. In language learning from text, h must be the index of a positive test procedure for the range of e. By suitable coding conventions, language learning from informant and logical learning can also modeled with correctness relations in the input stream paradigm. Computational analogs of the Borel complexity classes can be defined for correctness relations, in which case analogous characterization theorems hold for computable inquiry (Kelly 1996).

The moral of this discussion is that the problem of induction, or empirical underdeterination, comes in degrees corresponding to standard topological and computational complexity classes, which determine the objective sense in which reliable inquiry is possible.

A Foolish Consistency

A *consistent* learner never produces an output that is incorrect of every relevantly possible input stream extending the current input sequence. For non-computable learners, consistency makes a great deal of sense: why should someone who aims to find the truth say what has to be wrong? On the other hand, we have seen that formal relations can pose an "internal" problem of induction for computable learners. Since we do not require omniscience on the empirical side, why should we do so on the formal side when the underlying structure of the problem of induction is the same on both sides?

That raises an interesting question. Could insistence on computationally achievable consistency *preclude* computationally achievable empirical reliability? The answer is striking. One can construct an empirical proposition with the following properties. (1) The proposition is computably refutable with certainty. (2) Some computable, consistent method exists for the proposition (the method that always says "false" suffices since the proposition is never verified). But (3) no consistent, computable method of even a highly idealized, uncomputable kind[19] can even gradually decide the hypothesis. Thus, where traditional epistemology sees consistency as a *means* for finding the truth sooner, enforcing *achievable* consistency may prevent computable learners from finding truths they could otherwise have reliably found. So if the aim of inquiry is to find the truth, inconsistency may be an epistemic *obligation* (rather than a merely forgivable lapse) for computable agents. Such results exemplify the sharp difference in emphasis between computational learning theory and traditional, justificationist epistemology.[20]

[19] i.e., hyperarithmetically definable.

[20] Osherson and Weinstein (1986) contains many restrictiveness results carrying a similar moral. Also, see Osherson and Weinstein (1988).

Gambling with Success

Suppose that each learning problem comes equipped with an assignment of probabilities to empirical propositions. More precisely, suppose that the probability assignment is defined on the set of all *Borel propositions* (i.e., the least set that contains all the open (Σ_1) propositions and that is closed under countable union and complementation). A *probability assignment* on the Borel propositions is a function taking values in the unit interval that assigns unity to the vacuous proposition and that is *finitely additive* in the sense that the probability of a finite union of mutually incompatible Borel propositions is the sum of the probabilities of the propositions the union is taken over. *Countable additivity* extends finite additivity to countable, disjoint unions. While Kolmogorov's familiar mathematical theory of probability assumes countable additivity as a postulate, limiting relative frequencies do not satisfy it, and the usual foundations of Bayesian probability theory do not entail it (e.g., DeFinetti 1990; Savage 1972).

Say that an hypothesis is gradually decidable *with probability* r just in case there exists some empirical proposition of probability r over which the hypothesis is gradually decidable in the usual sense, and similarly for the other assessment criteria. Probabilistic success can be much easier to achieve than success in each relevant possibility. If the probability assignment is countably additive, then, remarkably, every Borel hypothesis is (1) decidable in the limit with unit probability and (2) decidable with certainty with arbitrarily high but non-unit probability. (1) can be improved to the result that the method of updating the given probability measure by conditionalization gradually decides the hypothesis with unit prior probability (e.g., Halmos 1974). That is a very general version of the familiar Bayesian claim that prior probabilities are eventually "swamped" by the input.

Compared with the purely topological analysis of section "Reliability and complexity", these probabilistic results sound too good to be true, since of error Borel propositions can be infinitely more complex than $\triangle_2$ propositions (Hinman 1978). What accounts for the dramatic difference? Suppose we want to decide the "zeros forever" hypothesis with a small, nonzero probability of error r. The negation of that hypothesis is the countable, disjoint union of the hypotheses $h_i =$ "the first nonzero occurs at position i". So by countable additivity, the probability that the "zeros forever" hypothesis is false is the sum of the probabilities of the propositions h_i. Since the infinite sum converges to a finite value, there is some position n such that the sum of the probabilities of h_n, h_{n+1}, ... is less than r. So our probability of failure is less than r if we halt with "true" at stage n if no nonzero datum has been seen by position n and halt with "false" as soon as a nonzero datum is seen. In other words, countable additivity asserts that when a high prior probability of successful learning suffices, only finitely many of the demon's opportunities to make the hypothesis false matter.

Without countable additivity, it is possible that the probability that the hypothesis is false exceeds the mass distributed over the h_n, say by a value of r. Since that "residual" probability mass is not distributed over the propositions h_i, the learner never "gets past" it, so whenever the learner halts inquiry with "true", the probability that the conclusion was in error remains at least as high as r. The residual probability

reflects the demon's *inexhaustible* opportunities to falsify the hypothesis in the infinite future, providing a probabilistic model of Sextus' demonic argument. In fact, both (1) and (2) can fail when countable additivity is dropped (Kelly 1996), highlighting the pivotal epistemological significance of that seemingly "technical" asumption.

Concept Learning and the PAC Paradigm

In the *Meno*, Plato outlined what has come to be known as the *concept learning paradigm*, which has captured the imagination of philosophers, psychologists, and artificial intelligence researchers ever since. A concept learning problem specifies a domain of *examples* described as vectors of *values* (e.g., blue, five kilos) of a corresponding set of *attributes* (e.g., color, weight), together with a set of possible *target concepts*, which are sets of examples. The learner is somehow presented with examples labelled either as positive or as negative examples of the concept to be learned, and the learner's task is to converge in some specified sense to a correct definition. In contemporary artificial intelligence and cognitive science, the "concepts" to be learned are defined by neural networks, logic circuits, and finite state automata, but the underlying paradigm would still be familiar to Socrates.

Socrates ridiculed interlocutors who proposed disjunctive concept definitions, which suggests that he would countenance only conjunctively definable concepts as relevant possibilities. Socrates' notorious solution to the concept learning problem was to have the environment "give away" the answer in a mystical flash of insight. But J. S. Mill's (i.e., Francis Bacon's) well-known inductive methods need no mystical help to identify conjunctive concepts with certainty: the first conjecture is the first positive example sampled. On each successive positive example in the sample, delete from the current conjecture each conjunct that disagrees with the corresponding attribute value of the example (the "method of difference"). On each successive negative example that agrees with the current conjecture everywhere except on one attribute, underline the value of that attribute in the current conjecture (the "method of similarity"). When all conjuncts in the current conjecture are underlined, halt inquiry.

Boolean concepts are also identifiable with certainty over a finite set of attribute values: wait for all possible examples to come in and then disjoin the positive ones. Bacon's methods sound plausible in the conjunctive case, whereas that "gerrymandering" procedure for learning Boolean concepts sounds hopeless (it is, in fact, just what Socrates ridiculed). Yet both procedures identify the truth with certainty, since the set of examples is finite. The PAC (Probably Approximately Correct) paradigm distinguishes such "small" problems in terms of *tractable* rather than merely *computable* inquiry.[21]

[21] An excellent source presenting all of the results mentioned here is Kearns and Vazirani (1994), which provides detailed descriptions and bibliographic notes for all the results mentioned below.

In the PAC paradigm, examples are sampled with replacement from an urn in which the probability of selecting an example is unknown. There is a collection of relevantly possible concepts and also a collection of hypotheses specifying the possible forms in which the learner is permitted to define a relevantly possible concept. Say that a hypothesis is ϵ-accurate just in case the sampling probability that a single sampled individual is a counterexample is less than ϵ. The learner is given a *confidence* parameter δ and an *error* parameter ϵ. From these parameters, the learner specifies a sample size and upon inspecting the resulting sample, she outputs a hypothesis. A learning strategy is *probably approximately correct* (PAC) just in case for each probability distribution on the urn and for each ϵ, δ exceeding zero, the strategy has a probability of at least $1-\epsilon$ of producing an ϵ-accurate hypothesis.

It remains to specify what it means for a PAC learning strategy to be *efficient*. Computational complexity is usually analyzed in terms of asymptotic growth *rate* over an infinite sequence of "similar" but "ever larger" examples of the problem. Tractability is understood as resource consumption bounded almost everywhere by some polynomial function of problem size. The size of a concept learning problem is determined by (1) the number of attributes (2) the size of the smallest definition of the target concept, (3) the reciprocal of the confidence parameter, and (4) the reciprocal of the error parameter (higher accuracy and reliability requirements make for a "bigger" inference problem). A *input efficient* PAC learner takes a sample in each problem whose size is bounded by a polynomial in these four arguments. Otherwise, the learner requires samples that are exponentially large in the same parameters, which is considered to be intractable.

There is an elegant combinatorial characterization of how large the sample required for PAC learning should be. Say that a concept class *shatters* a set S of examples just in case each subset of S is the intersection of S with some concept in the class. The *Vapnik-Chervonenkis* (VC) dimension of the concept class is the cardinality of the largest set of instances shattered by the class. There exists a fixed constant c such that if the VC dimension of the concept class is d, it suffices for PAC learnability that a sample of size s be taken, where:

$$s \geq c\left(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log\frac{1}{\epsilon}\right)$$

For example, the VC dimension of the conjunctive concepts over n Boolean attributes is $2n$ (and in fact is just n if $n>1$) so the problem is input-efficiently solvable by setting the sample size according to the above formula and then using any method producing conjectures consistent with the input (e.g., Bacon's method of similarity). Calculating the VC dimension of the concepts decidable by neural networks reveals that they are also input-efficiently learnable.

On the negative side, it can be shown that if the VC dimension of a concept class is d, then on some concept and in some sampling distribution, a sample size of at least d/ϵ is required. Since the VC dimension of the Boolean concepts over n Boolean attributes is 2^n, exponentially large samples will sometimes be required. Thus, any algorithm that takes a sample whose size depends only on the problem

and not the size of the (unknown) target concept itself will be input-inefficient (since the sample size grows non-polynomially when concept size is held fixed at the minimum value).

A *computationally efficient* PAC learner is a PAC learner whose runtime is bounded by a polynomial of the sort described in the definition of input efficiency. Since scanning a sampled instance takes time, computational efficiency implies input efficiency. Since Bacon's method is computationally trivial and requires small samples, it is a computationally efficient PAC learner. Bacon's method can be generalized to efficiently PAC learn k-CNF concepts (i.e., conjunctions of k-ary disjunctions of atomic or negated atomic sentences), for fixed k.

Sometimes computational difficulties arise entirely because it is hard for the learner to frame her conjecture in the required hypothesis language. It is known, for example, that the k-term DNF concepts (i.e., disjunctions of k purely conjunctive concepts) are not efficiently PAC learnable using k-term DNF hypotheses (when $k \geq 2$),[22] whereas they are efficiently PAC learnable using k-CNF hypotheses.

For some time it was not known whether there exist efficiently solvable PAC problems that are unsolvable neither due to sample-size complexity nor due to output representation. It turns out (Kearns and Valiant 1994) that under a standard cryptographic hypothesis,[23] the Boolean concepts of length polynomial in the number of attributes have that property, as does the neural network training problem.

An alternative way to obtain more refined results in a non-probabilistic context is to permit the learner to ask questions. A *membership oracle* accepts an example from the learner and returns "in" or "out" to indicate whether it is a positive or a negative example. A *Socratic oracle* responds to an input conjecture with a counterexample, if there is one.[24] One such result is that Socratic and membership queries suffice for identification of finite state automata with certainty in polynomial time (Angluin 1987).

Learning Theory and Epistemology

To coherentists, learning theory looks like a naive form of foundationalism, in which incorrigible beliefs are the fulcrum driving inquiry to the truth. But foundationalists are also disappointed because positive learning theoretic results depend on substantial, contingent assumptions such as the nature of the signals from the environment, the structure of time, and the range of relevant possibilities. Externalists would prefer to investigate *our* reliability directly, instead of taking a mathematical detour into possible methods and problems. And contextualists will object to the fixity

[22] This negative result holds only under the familiar complexity-theoretic hypothesis that $P \neq NP$.

[23] I.e, that computing discrete cube roots is intractable even for random algorithms.

[24] In the learning theoretic literature, Socratic queries are referred to as "equivalence" queries.

of truth through time, ignoring the possibility of meaning shifts due to conceptual change.

But on a more careful examination, learning theory amplifies some recent epistemological trends. The search for incorrigible foundations for knowledge is no longer considered a serious option, so the fact that reliability depends on contingent assumptions is hardly a penetrating objection. Indeed, it can be shown by learning theoretic means that if some background knowledge is necessary for reliability, that knowledge cannot be reliably assessed according to the same standard, blocking any attempt at an entirely reliability-based foundationalism.

Externalist epistemologies sidestep the foundational demand that the conditions for reliability be known by requiring only that we be reliable, without necessarily being aware of that fact. Knowledge attributions are then empirical hypotheses that can be studied by ordinary empirical means. But mature empirical investigations are always focused by general mathematical constraints on what is possible. Accordingly, learning theoretic results constrain naturalistic epistemology by specifying how reliable an arbitrary system, whether computable or otherwise, could possibly be in various learning situations.

Externalism has encountered the objection (Lehrer 1990) that reliability is insufficient for knowledge if one is not justified in believing that one is reliable (e.g., someone has a thermometer implanted in her brain that suddenly begins to produce true beliefs about the local temperature). The intended point of such objections is that reliable belief-forming processes should be embedded in a coherent belief system incorporating beliefs about the agent's own situation and reliability therein. Learning theory may then be viewed as defining the crucial relation of *methodological coherence* between epistemic situations, ambitions, and means. Unlearnability arguments isolate methodological incoherence and positive arguments suggest methods, background assumptions, or compromised ambitions which, if adopted, could bring a system of beliefs into methodological coherence.

Incorporating learning theoretic structure into the concept of coherence addresses what some coherentists take to be the chief objection to their position.

> ... [A]lthough any adequate epistemological theory must confront the task of bridging the gap between justification and truth, the adoption of a nonstandard conception of truth, such as a coherence theory of truth, will do no good unless that conception is independently motivated. Therefore, it seems that a coherence theory of justification has no acceptable way of establishing the essential connection with truth (Bonjour 1985): 110.

Whether a methodological principle guarantees or prevents reliable convergence to the truth is, of course, the unshakable focus of learning theoretic analysis. Where coherence is at issue, one must consider a multitude of possible interpretations of reliability and of one's epistemic situation, backing and filling until the analysis seems apt and fits with the rest of one's beliefs. That pluralistic attitude is reflected in the wide variety of success criteria, paradigms and problems considered in the learning theoretic literature.

Contextualists may also find some value in learning theoretic results. The first moral of the subject is that reliability is highly sensitive to the finest details of the input presentation, the range of possible alternatives, the kinds of hypotheses or skills at issue, the learner's cognitive powers and resources, and the methodological

principles to which she is committed. Reliable methodology is unavoidably piecemeal, contextual methodology, optimized to the special features of the problem at hand.

A remaining contextualist objection is that learning theory presupposes a fixed "conceptual scheme" in which truth is a fixed target, whereas in light of conceptual revolutions, meaning and hence truth changes as the beliefs of the learner change through time. That objection does apply to the usual learning theoretic paradigms, but the concept of reliability is flexible enough to accommodate it. If truth feints as inquiry lunges, then success can be defined as a methodological *fixed point* in which the beliefs of the learner are eventually true *with respect to themselves* (Kelly 1996; Kelly and Glymour 1992). Unlike norms of justification, which may change through time, convergence to the *relative* truth provides a strategic aim that plausibly survives successive changes in the underlying scientific tradition.

Bibliography

Angluin, D. (1980). Inductive inference of formal languages from positive data. *Information and Control, 45*(2), 117–135.

Angluin, D. (1987). Learning regular sets from queries and counterexamples. *Information and Computation, 75*, 87–106.

Blum, M., & Blum, L. (1975). Toward a mathematical theory of inductive inference. *Information and Control, 28*, 125–155.

Bonjour, L. (1985). *The structure of empirical knowledge*. Cambridge: Harvard University Press.

Brown, R., & Hanlon, C. (1970). Derivational complexity and the order of acquisition of child speech. In J. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley.

Carnap, R. (1950). *The logical foundations of probability*. Chicago: University of Chicago Press.

Case, J., & Smith, C. (1983). Comparison of identification criteria for machine inductive inference. *Theoretical Computer Science, 24*, 193–220.

DeFinetti. (1990). *The theory of probability*. New York: Wiley.

Glymour, C. (1980). *Theory and evidence*. Cambridge: M.I.T. Press.

Gold, E. M. (1965). Limiting recursion. *Journal of Symbolic Logic, 30*, 27–48.

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*, 447–474.

Halmos, P. (1974). *Measure theory*. New York: Springer.

Hinman, P. (1978). *Recursion theoretic hierarchies*. New York: Springer.

James, W. (1948). The will to believe. In A. Castell (Ed.), *Essays in pragmatism*. New York: Collier Macmillan.

Kearns, M., & Valiant, L. (1994). Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM, 41*, 57–95.

Kearns, M., & Vazirani, U. (1994). *An introduction to computational learning theory*. Cambridge: M.I.T. Press.

Kelly, K. (1992). Learning theory and descriptive set theory. *Logic and Computation, 3*, 27–45.

Kelly, K. (1996). *The logic of reliable inquiry*. New York: Oxford University Press.

Kelly, K., & Glymour, C. (1989). Convergence to the truth and nothing but the truth. *Philosophy of Science, 56*, 185–220.

Kelly, K., & Glymour, C. (1990). Theory discovery from data with mixed quantifiers. *Journal of Philosophical Logic, 19*, 1–33.

Kelly, K., & Glymour, C. (1992). Inductive inference from theory-laden data. *Journal of Philosophical Logic, 21*, 391–444.

Kelly, K., & Schulte, O. (1995). The computable testability of theories making uncomputable predictions. *Erkenntnis, 43*, 29–66.
Kelly, K., & Schulte, O. (1997). Church's thesis and Hume's problem. In M. L. Dalla Chiara et al. (Eds.), *Logic and scientific methods*. Dordrecht: Kluwer.
Kemeny, J. (1953). The use of simplicity in induction. *Philosophical Review, 62*, 391–408.
Kugel, P. (1977). Induction pure and simple. *Information and Control, 33*, 236–336.
Lauth, B. (1993). Inductive inference in the limit for first-order sentences. *Studia Logica, 52*, 491–517.
Lehrer, K. (1990). *Theory of knowledge*. San Francisco: Westview.
Levi, I. (1991). *The fixation of belief and its undoing*. Cambridge: Cambridge University Press.
Miller, D. (1974). On Popper's definitions of verisimilitude. *British Journal of the Philosophy of Science, 25*, 155–188.
Mormann, T. (1988). Are all false theories equally false? *British Journal for the Philosophy of Science, 39*, 505–519.
Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, 231*(A), 289–337.
Osherson, D., & Weinstein, S. (1986). *Systems that learn*. Cambridge: M.I.T. Press.
Osherson, D., & Weinstein, S. (1988). Mechanical learners pay a price for Bayesianism. *Journal of Symbolic Logic, 56*, 661–672.
Osherson, D., & Weinstein, S. (1989a). Paradigms of truth detection. *Journal of Philosophical Logic, 18*, 1–41.
Osherson, D., & Weinstein, S. (1989b). Identification in the limit of first order structures. *Journal of Philosophical Logic, 15*, 55–81.
Osherson, D., & Weinstein, S. (1991). A universal inductive inference machine. *Journal of Symbolic Logic, 56*, 661–672.
Popper, K. (1968). *The logic of scientific discovery*. New York: Harper.
Popper, K. (1982). *Unended quest: An intellectual autobiography*. LaSalle: Open Court.
Putnam, H. (1963). Degree of confirmation' and inductive logic. In A. Schilpp (Ed.), *The philosophy of Rudolph Carnap*. LaSalle: Open Court.
Putnam, H. (1965). Trial and error predicates and a solution to a problem of Mostowski. *Journal of Symbolic Logic, 30*, 49–57.
Reichenbach, H. (1938). *Experience and prediction*. Chicago: University of Chicago Press.
Savage, L. (1972). *The foundations of statistics*. New York: Dover.
Sextus Empiricus. (1985). *Selections from the major writings on scepticism, man and god* (P. Hallie, Ed., S. Etheridge, Trans.). Indianapolis: Hackett.
Shapiro, E. (1981). *Inductive inference of theories from facts*. Report YLU 192. New Haven: Department of Computer Science, Yale University.
Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition*. Cambridge: M.I.T. Press.

Chapter 34
Some Computational Constraints in Epistemic Logic

Timothy Williamson

Introduction

This paper concerns limits that some epistemic logics impose on the complexity of an epistemic agent's reasoning, rather than limits on the complexity of the epistemic logic itself.

As an epistemic agent, one theorizes about a world which contains the theorizing of epistemic agents, including oneself. Epistemic logicians theorize about the abstract structure of epistemic agents' theorizing. This paper concerns the comparatively simple special case of epistemic logic in which only one agent is considered. Such an epistemic agent theorizes about a world which contains that agent's theorizing. One has knowledge about one's own knowledge, or beliefs about one's own beliefs. The considerations of this paper can be generalized to multi-agent epistemic logic, but that will not be done here. Formally, single-agent epistemic logic is just standard monomodal logic; we call it 'epistemic' in view of the envisaged applications.

In epistemic logic, we typically abstract away from some practical computational limitations of all real epistemic agents. For example, we are not concerned with their failure to infer from a proposition q the disjunction $q \vee r$ for every unrelated proposition r. What matters is that if some propositions do in fact follow from the agent's theory (from what the agent knows, or believes), then so too do all their logical consequences. For ease of exposition, we may idealize epistemic agents and describe them as knowing whatever follows from what they know, or as believing whatever follows from what they believe, but we could equally well redescribe the matter in less contentious terms by substituting 'p follows from what one knows'

T. Williamson (✉)
University of Oxford, Oxford, UK
e-mail: timothy.williamson@philosophy.ox.ac.uk

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_34

for 'one knows p' or 'p follows from what one believes' for 'one believes p' throughout the informal renderings of formulas, at the cost only of some clumsiness. Thus, if we so wish, we can make what looks like the notorious assumption of logical omniscience true by definition of the relevant epistemic operators. On suitable readings, it is a triviality rather than an idealization. It does not follow that no computational constraints are of any concern to epistemic logic. For if one's knowledge is logically closed by definition, that makes it computationally all the harder to know that one does *not* know something: in the standard jargon, logical omniscience poses a new threat to negative introspection. That threat is one of the phenomena to be investigated in this paper.

In a recursively axiomatizable epistemic logic, logical omniscience amounts to closure under a recursively axiomatizable system of inferences. Thus all the inferences in question can in principle be carried out by a single Turing machine, an idealized computer. Epistemic logicians do not usually want to make assumptions which would require an epistemic agent to exceed every Turing machine in computational power. In particular, such a requirement would presumably defeat the purpose of the many current applications of epistemic logic in computer science. By extension, epistemic logicians might prefer not to make assumptions which would permit an epistemic agent not to exceed every Turing machine in computational power only under highly restrictive conditions. Of course, such assumptions might be perfectly appropriate in special applications of epistemic logic to cases in which those restrictive conditions may be treated as met. But they would not be appropriate in more general theoretical uses of epistemic logic.

As an example, let us consider the so-called axiom of *negative introspection* alluded to above. It may be read as the claim that if one does not know p then one knows that one does not know p, or that if one does not believe p then one believes that one does not believe p. In terms of theories: if one's theory does not entail p, then one's theory entails that one's theory does not entail p. That assumption is acceptable in special cases for special values of 'p'. However, for a theory to be consistent is in effect for there to be some p which it does not entail. On this reading, negative introspection implies that if one's theory is consistent then it entails its own consistency. But, by Gödel's second incompleteness theorem, if one's theory is recursively axiomatizable and includes Peano arithmetic, then it entails its own consistency only if it is inconsistent. Thus, combined with the incompleteness theorem, negative introspection implies that if one's theory is recursively axiomatizable then it includes Peano arithmetic only if it is inconsistent. Yet, in a wide range of interesting cases, the output of a Turing machine, or the theory of an epistemic agent of equal computational power, is a consistent recursively axiomatizable theory which includes Peano arithmetic. Thus, except in special circumstances, the negative introspection axiom imposes an unwarranted constraint on the computational power of epistemic agents.

Naturally, such an argument must be made more rigorous before we can place much confidence in it. That will be done below. The problem for the negative introspection axiom turns out to be rather general: it arises not just for extensions of Peano arithmetic but for any undecidable recursively axiomatizable theory, that is,

for any theory which is the output of some Turing machine while its complement is not. It is very natural to consider epistemic agents whose theories are of that kind.

The aim of this paper is not primarily to criticize the negative introspection axiom. Rather, it is to generalize the problem to which that axiom gives rise, to formulate precisely the conditions which a system of epistemic logic must satisfy in order not to be susceptible to such problems, and to investigate which systems satisfy those conditions. The conditions in question will be called *r.e. conservativeness* and *r.e. quasi-conservativeness*. Very roughly indeed, a system satisfies these conditions if it has a wide enough variety of models in which the epistemic agent is computationally constrained. Such models appear to be among the intended models on various applications of epistemic logic. As already noted, systems of epistemic logic which do not satisfy the conditions may be appropriate for other applications. But it is time to be more precise.

Elementary Epistemic Logic

Let L be the language consisting of countably many propositional variables p_0, p_1, p_2, ... (p and q represent arbitrary distinct variables), the falsity constant $\perp$ and the material conditional $\supset$. Other operators are treated as metalinguistic abbreviations in the usual way. We expand L to the language $L_\Box$ of propositional modal logic by adding the operator $\Box$. $\Diamond\alpha$ abbreviates $\neg\Box\neg\alpha$. Unless otherwise specified, the metalinguistic variables α, β, γ, ... range over all formulas of $L_\Box$. We use the necessity symbol $\Box$ from modal logic to make various formulas and formal systems look familiar, without prejudice to its interpretation. We reinterpret $\Box$ as something like 'I know that' or 'I believe that'. To generalize over reinterpretations, we use the neutral verb 'cognize' for $\Box$ in informal renditions of formulas.

A *theory* in $L_\Box$ is a subset of $L_\Box$ containing all truth-functional tautologies and closed under modus ponens for $\supset$ (MP). A *model* M of $L_\Box$ induces a function $M() : L_\Box \to \{0, 1\}$ where $M(\perp) = 0$ and $M(\alpha \supset \beta) = 1$ if and only if $M(\alpha) \leq M(\beta)$. Intuitively, $M(\alpha) = 1$ if and only if α is true in M; $M(\alpha) = 0$ if and only if α is false in M. An *application* of epistemic logic determines a class of its intended models. The logic of the application is the set of formulas α such that $M(\alpha) = 1$ for every intended model M; thus the logic is a theory in $L_\Box$. Of course, we can also define a relation of logical consequence on the models, but for present purposes it is simpler to identify a logic with the set of its theorems.

Since atomic sentences are treated simply as propositional variables, we may substitute complex formulas for them. More precisely, we assume that for each intended model M and uniform substitution σ there is an intended model M^σ such that for every α $M^\sigma(\alpha) = M(\sigma\alpha)$. Thus the logic of the application is closed under uniform substitution (US).

A *modal logic* is a theory in $L_\Box$ closed under US. The logic of an application is a modal logic. The smallest modal logic is PC, the set of all truth-functional

tautologies. If Σ is a modal logic, we write $\vdash_\Sigma \alpha$ when $\alpha \in \Sigma$. For any $X \subseteq L_\Box$, we define $X \vdash_\Sigma \alpha$ if and only if $\vdash_\Sigma \wedge X_0 \supset \alpha$ for some finite $X_0 \subseteq X$ ($\wedge X_0$ and $\vee X_0$ are the conjunction and disjunction respectively of X_0 on a fixed ordering of the language). X is Σ-consistent unless $X \vdash_\Sigma \bot$. A maximal Σ-consistent set is a Σ-consistent set not properly included in any Σ-consistent set.

If M is a model, let $\Box^{-1}M = \{\alpha : M(\Box\alpha) = 1\}$. Thus $\Box^{-1}M$ expresses what the agent cognizes in M. If Σ is the logic of an application on which $\Box^{-1}M$ is a theory in $L_\Box$ for every intended model M, then for all formulas α and β, $\vdash_\Sigma \Box(\alpha \supset \beta) \supset (\Box\alpha \supset \Box\beta)$ (axiom schema K) and if $\vdash_{PC} \alpha$ then $\vdash_\Sigma \Box\alpha$ (rule RN_{PC}). A modal logic satisfying RN_{PC} and K is *prenormal*. If cognizing is knowing or believing, then prenormality is an extreme idealization, a form of logical omniscience. But if cognizing is the closure of knowing or believing under at least truth-functional consequence, then prenormality is innocuous. The rule RN_{PC} differs from the stronger and better-known rule RN (necessitation or epistemization): if $\vdash_\Sigma \alpha$ then $\vdash_\Sigma \Box\alpha$. A modal logic Σ satisfying RN and K is *normal*. Unlike RN_{PC}, RN requires the agent to cognize all theorems of the epistemic logic, not just all truth-functional tautologies. For instance, $\Box\top$ is a theorem of every prenormal logic by RN_{PC}, but since it is not a theorem of PC we cannot iterate the rule; $\Box\Box\top$ is not a theorem of the smallest prenormal logic. By contrast, we can iterate RN, and $\Box\Box\top$ is a theorem of every normal modal logic. Prenormality does not imply that agents cognize their own cognizing. It merely implies that they can formulate propositions about cognizing, for since $\Box\alpha \supset \Box\alpha$ is a truth-functional tautology, $\Box(\Box\alpha \supset \Box\alpha)$ is a theorem of every prenormal logic. Since normality entails prenormality, results about all prenormal logics apply to all normal modal logics. Every logic extending a prenormal logic is prenormal; by contrast, some nonnormal logics extend normal logics, although any extension of a normal logic is at least prenormal.

Any normal logic Σ has a possible worlds semantics where $\Box\alpha$ is true at a world w in a model M if and only if α is true at every world in M to which w has the accessibility relation of M. Intuitively, a world x is accessible from w if and only if what the agent cognizes at w is true at x. In other words, one world is accessible from another if and only if for all one cognizes in the latter one is in the former. The formulas α such that $\Box\alpha$ is true at w express what the agent cognizes at w. For every normal logic Σ there is a class C of models such that Σ consists of exactly the formulas true at every world in every model in C.

Many authors require the accessibility relation to be an equivalence relation (reflexive, symmetric and transitive) for every intended model of their application. A common attitude is expressed by the authors of a standard text, who write that the postulate 'seems reasonable for many applications we have in mind', but 'we can certainly imagine other possibilities' (Fagin et al. 1995, 33). For example, if x is accessible from w if and only if appearances to the agent are identical in x and w, then accessibility is an equivalence relation because identity in any given respect is an equivalence relation. The logic of the class of all possible worlds models in which accessibility is an equivalence relation is the modal system known as S5 : $\vdash_{S5} \alpha$ if and only if α is true in every model for which accessibility is an equivalence relation. Since equivalence relations correspond to partitions of the set

of worlds, S5 is also known as the logic of the *partitional* conception of knowledge. S5 is the smallest normal modal logic with the theorem schemas T($\Box\alpha \supset \alpha$) and E ($\neg\Box\alpha \supset \Box\neg\Box\alpha$). T (*truthfulness*) says that the agent cognizes only truths; it is appropriate for applications on which one cognizes only what follows from what one knows. T corresponds to the condition that accessibility be reflexive. For applications on which one cognizes what follows from what one *believes*, T would have to be dropped, perhaps replaced by the weaker principle D ($\Box\alpha \supset \Diamond\alpha$). D requires cognition to be consistent in the sense that an agent who cognizes something does not also cognize its negation. D corresponds to the condition that accessibility be serial (from every world some world is accessible). E is the principle of *negative introspection*: cognition records its omissions in the sense that agents who do not cognize something cognize that they do not cognize it. E corresponds to the condition that accessibility be euclidean (worlds accessible from a given world are accessible from each other). In S5 we can derive the principle of *positive introspection* 4 ($\Box\alpha \supset \Box\Box\alpha$), that cognition records its contents in the sense that agents who cognize something cognize that they cognize it. 4 corresponds to the condition that accessibility be transitive. If T is dropped or weakened to D then 4 is no longer derivable from E, so 4 might be added as an independent schema. Accessibility is reflexive (T) and euclidean (E) if and only if it is an equivalence relation.

Computational Constraints

To formulate computational constraints, we generalize concepts from recursion theory to $L_\Box$ using a standard intuitively computable coding procedure. A model M is r.e. if and only if $\Box^{-1}M$ (which expresses what the agent cognizes in M) is an r.e. (recursively enumerable) theory in $L_\Box$. In that sense, the agent's cognition in an r.e. model does not exceed the computational capacity of a sufficiently powerful Turing machine.

Consider the restriction of $\Box^{-1}M$ to the $\Box$-free sublanguage L, $L \cap \Box^{-1}M$. Let $\Box^{-1}M$ be an r.e. theory in $L_\Box$. Thus $L \cap \Box^{-1}M$ is an r.e. theory in L. It is the part of the agent's overall theory in M which is not specifically epistemic. From the standpoint of general epistemic logic, can we reasonably impose any further constraints on $L \cap \Box^{-1}M$ beyond recursive enumerability?

If $\Box^{-1}M$ is required to be consistent, $L \cap \Box^{-1}M$ is consistent too. Can we limit the possible values of $L \cap \Box^{-1}M$ still further? For many applications we cannot. $L \cap \Box^{-1}M$ simply expresses what the agent cognizes in M about some aspect of reality. The agent can store any r.e. theory in L as a recursive axiomatization (Craig 1953). If the agent might cognize that aspect of reality simply by having learned a theory about it on the testimony of a teacher, any (consistent) r.e. theory in L is possible. In particular, we can interpret the propositional variables as mutually independent. For example, given a black box which may or may not flash a light on input of a symbol for a natural number, we can read p_i as 'The light flashes on input *i*'. Then any (consistent) r.e. theory in L could exhaust everything expressible

in L which the agent (with only the computational power of a Turing machine) has learned about the black box. Such situations seem quite reasonable. If we want an epistemic logic to have a generality beyond some local application, it should apply to them: such situations should correspond to intended models. Now any application which has all those intended models thereby satisfies (*) or ($*_{con}$), depending on whether the epistemic agent's theory is required to be consistent:

(*) For every r.e. theory R in L, $L \cap \Box^{-1}M = R$ for some r.e. intended model M.

($*_{con}$) For every consistent r.e. theory R in L, $L \cap \Box^{-1}M = R$ for some r.e. intended model M.

(*) is appropriate for readings of $\Box$ like 'It follows from what I believe that ...', if the agent is not required to be consistent. For readings of $\Box$ like 'It follows from what I know that ...', only ($*_{con}$) is appropriate, for one can know only truths and any set of truths is consistent. We can define corresponding constraints on a modal logic Σ without reference to models:

Σ is *r.e. conservative* if and only if for every r.e. theory R in L, there is a maximal Σ-consistent set X such that $\Box^{-1}X$ is r.e. and $L \cap \Box^{-1}X = R$.

Σ is *r.e. quasi-conservative* if and only if for every consistent r.e. theory R in L, there is a maximal Σ-consistent set X such that $\Box^{-1}X$ is r.e. and $L \cap \Box^{-1}X = R$.

Here $\Box^{-1}X = \{\alpha \in L_{\Box} : \Box\alpha \in X\}$. Roughly, if Σ is r.e. (quasi-)conservative then every (consistent) r.e. theory in the language without $\Box$ is conservatively extended by an r.e. theory in the language with $\Box$ such that it is consistent in Σ for R to be exactly what the agent cognizes in the language without $\Box$ while what the agent cognizes in the language with $\Box$ constitutes an r.e. theory. If an application satisfies (*), its logic is r.e. conservative, for X can be the set of formulas true in M. Conversely, any r.e. conservative logic is the logic of some application which satisfies (*), for some appropriate kind of intended model. The same relationships hold between ($*_{con}$) and r.e. quasi-conservativeness. For many applications of epistemic logic, the class of intended models is quite restricted and even ($*_{con}$) does not hold. But if the application interprets $\Box$ as something like 'It follows from what I believe/know that', without special restrictions on the epistemic subject, then situations of the kind described above will correspond to intended models and the logic of the application will be r.e. [quasi-] conservative. In this paper we do not attempt to determine which informally presented applications of epistemic logic satisfy (*) or ($*_{con}$). We simply investigate which logics are r.e. [quasi-] conservative.

Trivially, every r.e. conservative modal logic is r.e. quasi-conservative. Examples will be given below of r.e. quasi-conservative normal modal logics which are not r.e. conservative. For prenormal modal logics, r.e. conservativeness can be characterized in terms of r.e. quasi-conservativeness in a simple way which allows us to transfer results about one to the other:

Proposition 34.1 Let Σ be a prenormal modal logic. Then Σ is r.e. conservative if and only if Σ is r.e. quasi-conservative and not $\vdash_{\Sigma} \Diamond\top$.

Proof Let $\Box\,\mathrm{L} = \{\Box\alpha : \alpha \in \mathrm{L}\}$. ($\Rightarrow$) Trivially, Σ is r.e. quasi-conservative if r.e. conservative. Suppose that $\vdash_\Sigma \Diamond\top$. Since $\Box\,\mathrm{L} \vdash_\Sigma \Box\neg\top$, $\Box\,\mathrm{L}$ is Σ-inconsistent. Thus $\mathrm{L} \cap \Box^{-1}\mathrm{X} = \mathrm{L}$ for no Σ-consistent set X. Since L is an r.e. theory in L, Σ is not r.e. conservative. ($\Leftarrow$) Suppose that Σ is r.e. quasi-conservative but not r.e. conservative. Since L is the only inconsistent theory in L, there is no maximal Σ-consistent set X such that $\Box^{-1}\mathrm{X}$ is r.e. and $\mathrm{L} \cap \Box^{-1}\mathrm{X} = \mathrm{L}$. If $\Box\,\mathrm{L}$ is Σ-consistent, then some maximal Σ-consistent set X extends $\Box\,\mathrm{L}$, so $\mathrm{L} \cap \Box^{-1}\mathrm{X} = \mathrm{L}$; but for $\alpha \in \mathrm{L}$ $\Box \vdash_\Sigma \bot \supset \alpha$, so $\vdash_\Sigma \Box \bot \supset \Box\alpha$ by prenormality, so $\Box^{-1}\mathrm{X} = \mathrm{L}_\Box$ because $\Box \bot \in X$, so $\Box^{-1}\mathrm{X}$ is r.e. Thus $\Box\,\mathrm{L}$ is Σ-inconsistent, i.e., for some $\alpha_0, \ldots, \alpha_m \in \mathrm{L}$, $\vdash_\Sigma \neg\wedge\{\Box\alpha_i : i \leq m\}$. But for $i \leq m$, $\vdash_\Sigma \Box\neg\top \supset \Box\alpha_i$ by prenormality, so $\vdash_\Sigma \neg\Box\neg\top$.

Examination of the proof shows that the prenormality condition can be weakened to this: if $\vdash_{\mathrm{PC}} \alpha \supset \beta$ then $\vdash_\Sigma \Box\alpha \supset \Box\beta$. An example of a reading of $\Box$ which verifies this weaker condition but falsifies prenormality is 'There is a subjective probability of at least x that', where $0 < x < 1$, for prenormality implies that $\vdash_\Sigma (\Box p \wedge \Box q) \supset \Box(p \wedge q)$, whereas this reading invalidates that formula. Prenormality can be weakened in similar ways for subsequent propositions.

R.e. conservativeness and r.e. quasi-conservativeness do not state upper or lower bounds on the epistemic agent's computational capacity. Rather, they state upper bounds on the strength of the epistemic logic itself; evidently a modal logic with an r.e. [quasi-] conservative extension is itself r.e. [quasi-] conservative. But too strong a logic can impose unwarranted restrictions on the agent's theory of the world given an upper bound on the agent's computational capacity.

Some Non-R.e. Quasi-Conservative Logics

Which modal logics are not r.e. [quasi-] conservative? Obviously, since $\vdash_{S5} \Diamond\top$, the logic S5 is not r.e. conservative. Since S5 is decidable, this does not result from non-recursiveness in S5 itself. More significantly:

Proposition 34.2 S5 is not r.e. quasi-conservative.

Proof (Skyrms 1978, 377 and Shin and Williamson 1994, Proposition 34.1 have similar proofs of related facts about S5): Let R be a non-recursive r.e. theory in L; R is consistent. Suppose that $\Box^{-1}\mathrm{X}$ is r.e. and $\mathrm{L} \cap \Box^{-1}\mathrm{X} = \mathrm{R}$ for some maximal S5-consistent set X. Now $\mathrm{L} - \mathrm{R} = \{\alpha : \Box\neg\Box\alpha \in \mathrm{X}\} \cap \mathrm{L}$. For if $\alpha \in \mathrm{L} - \mathrm{R}$ then $\Box\alpha \notin \mathrm{X}$, so $\neg\Box\alpha \in \mathrm{X}$; but $\vdash_{S5} \neg\Box\alpha \supset \Box\neg\Box\alpha$, so $\Box\neg\Box\alpha \in \mathrm{X}$ since X is maximal S5-consistent. Conversely, if $\Box\neg\Box\alpha \in \mathrm{X}$ then $\neg\Box\alpha \in \mathrm{X}$ since $\vdash_{S5}$ $\Box\neg\Box\alpha \supset \neg\Box\alpha$, so $\Box\alpha \notin \mathrm{X}$, so $\alpha \notin \mathrm{R}$ since $\mathrm{L} \cap \Box^{-1}\mathrm{X} = \mathrm{R}$. Since $\Box^{-1}\mathrm{X}$ is r.e., so is $\{\alpha : \Box\neg\Box\alpha \in \mathrm{X}) \cap \mathrm{L}$, i.e. L – R. Contradiction.

Thus the partitional conception of knowledge prevents a subject with the computational capacity of a Turing machine from having as the restriction of its theory to the $\Box$-free language any non-recursive r.e. theory (for other problems with the S5 schema in epistemic logic and further references see Williamson (2000, 23–24, 166–167, 226–228, 316–317)). Thus S5 is unsuitable as a general epistemic logic for Turing machines.

The proof of Proposition 34.2 depends on the existence of an r.e. set whose complement is not r.e. By contrast, the complement of any recursive set is itself recursive; decidability, unlike semi-decidability, is symmetric between positive and negative answers. The analogue of Proposition 34.2 for a notion like r.e. quasi-conservativeness but defined in terms of recursiveness rather than recursive enumerability would be false. For it is not hard to show that if R is a consistent recursive theory in L, then there is a maximal S5-consistent set X in $L_\Box$ such that $\Box^{-1}X$ is recursive and $L \cap \Box^{-1}X = R$. Thus S5 imposes computational constraints not on very clever agents (whose theories need not be r.e.) or on very stupid agents (whose theories must be recursive) but on half-clever agents (whose theories must be r.e. but need not be recursive).

Proposition 34.2 is the rigorous version of the argument sketched in the introduction. Can we generalize it? The next result provides a rather unintuitive necessary condition for r.e. quasi-conservativeness which nevertheless has many applications.

Theorem 34.3 Let Σ be a modal logic such that for some formulas $\alpha_0, \ldots, \alpha_n \in L_\Box$ and $\beta_0, \ldots, \beta_n \in L$, $\vdash_\Sigma \vee\{\Box\alpha_i : i \leq n\}$ and, for each $i \leq n$, $\vdash_\Sigma (\Box\alpha_i \wedge \Box\beta_i) \supset \Box\bot$ and not $\vdash_{PC} \neg\beta_i$. Then Σ is not r.e. quasi-conservative.

Proof There are pairwise disjoint r.e. subsets $I_0, I_1, I_2, \ldots$ of the natural numbers $\mathbb{N}$ such that for every total recursive function f, $i \in I_{f(i)}$ for some $i \in \mathbb{N}$. For let $f[0]$, $f[1], f[2], \ldots$ be a recursive enumeration of all partial and total recursive functions on $\mathbb{N}$ and set $I_i = \{j : f[j](j)$ is defined and $= i\}$; then $j \in I_{f[j](i)}$ whenever $f[j]$ is total, I_i is r.e. and $I_i \cap I_j = \{\}$ whenever $i \neq j$. Now suppose that (i) $\vdash_\Sigma \vee\{\Box\alpha_i : i \leq n\}$; (ii) $\vdash_\Sigma (\Box\alpha_i \wedge \Box\beta_i) \supset \Box\bot$ for each $i \leq n$; (iii) $\vdash_{PC} \neg\beta_i$ for no $i \leq n$. Let m be the highest subscript on any propositional variable occurring in $\beta_0, \ldots, \beta_n$. For all $i \in \mathbb{N}$, let σ_i and τ_i be substitutions such that $\sigma_i\, p_j = p_{i(m+1)+j}$ and $\tau_i\, p_{i(m+1)+j} = p_j$ for all $j \in \mathbb{N}$. Set $U = \{\sigma_i\beta_j : i \in I_j\}$. Since the σ_i are recursive and the I_j are r.e., U is r.e. Now $\vdash_{PC} \neg\sigma_i\beta_j$ for no i, j, otherwise $\vdash_{PC} \neg\tau_i\sigma_i\beta_j$, i.e., $\vdash_{PC} \neg\beta_j$, contrary to (iii). Moreover, if $h \neq i$ then $\sigma_h\beta_j$ and $\sigma_i\beta_k$ have no propositional variable in common. Thus if $h \in I_j$ and $i \in I_k$ and $\sigma_h\beta_j$ has a variable in common with $\sigma_i\beta_k$, then $h = i$, so $j = k$ because the I_j are pairwise disjoint. Hence no two members of U have a propositional variable in common. Thus U is consistent. Let R be the smallest theory in L containing U; R is consistent and r.e. Suppose that for some maximal Σ-consistent set X, $\Box^{-1}X$ is r.e. and $L \cap \Box^{-1}X = R$. Let the total recursive function g enumerate $\Box^{-1}X$. Fix $j \in \mathbb{N}$. By (i), $\vdash_\Sigma \vee\{\Box\sigma_j\alpha_i : i \leq n.\}$ since Σ is closed under US, so $\Box\sigma_j\alpha_i \in Y$ for some $i \leq n$ since Y is maximal Σ-consistent. Thus $g(k) = \sigma_j\alpha_i$ for some k; let $k(j)$ be the least k such that $g(k) \in \{\sigma_j\alpha_i : i \leq n\}$. Let $f(j)$ be the least $i \leq n$ such that $g(k(j)) = \sigma_j\alpha_i$. Since g enumerates $\Box^{-1}X$, $\Box\sigma_j\alpha_{f(j)} \in X$. Since g and σ_j are total recursive, k is total recursive, so f is total recursive. Thus $j \in I_{f(j)}$ for some $j \in \mathbb{N}$, so $\sigma_j\beta_{f(j)} \in U \subseteq R$ since $f(j) \leq n$. Since $L \cap \Box^{-1}X = R$, $\Box\sigma_j\beta_{f(j)} \in X$. By (ii), $\vdash_\Sigma (\Box\alpha_{f(j)} \wedge \Box\beta_{f(j)}) \supset \Box\bot$, so $\vdash_\Sigma (\Box\sigma_j\alpha_{f(j)} \wedge \Box\sigma_j\beta_{f(j)}) \supset \Box\bot$; since X is maximal Σ-consistent, $\Box\bot \in X$. Thus $\bot \in R$, contradicting the consistency of R. Thus no such set as X can exist, so Σ is not r.e. quasi-conservative.

In other words, a necessary condition for Σ to be r.e. quasi-conservative is that for all formulas $\alpha_0, \ldots, a_n \in L_\Box$ and $\beta_0, \ldots, \beta_n \in L$, if $\vdash_\Sigma \vee(\Box\alpha_j : i \leq n)$ and, for each $i \leq n$, $\vdash_\Sigma(\Box\alpha_i \wedge \Box\beta_i) \supset \Box\bot$ then, for some $i \leq n$, $\vdash_{PC} \neg\beta_i$. Of course, if Σ is prenormal and contains the D axiom (requiring the agent to be consistent) then the condition that $\vdash_\Sigma(\Box\alpha_i \wedge \Box\beta_i) \supset \Box\bot$ can be simplified to the condition that $\vdash_\Sigma \neg\Box(\alpha_i \wedge \beta_i)$.

Open Problem Is the necessary condition for r.e. quasi-conservativeness in Theorem 34.3 (or some natural generalization of it) also sufficient?

Observation The proof of Theorem 34.3 uses significantly more recursion theory than does the proof of Proposition 34.2, which relies only on the existence of an r.e. set whose complement is not r.e. Samson Abramsky observed (informal communication) that the proof of Proposition 34.2 would generalize to a setting in which r.e. sets were replaced by open sets in a suitable topology (in which not all open sets have open complements). It would be interesting to see whether a generalization along such lines yielded a smoother theory. One might then seek an intuitive interpretation of the topology.

To see that Proposition 34.2 is a special case of Theorem 34.3, put $n = 1$, $\alpha_0 = \Diamond\neg p$, $\alpha_1 = \Diamond p$, $\beta_0 = p$ and $\beta_1 = \neg p$. Now; $\vdash_{S5} \Box\Diamond\neg p \vee \Box\Box p$; $\vdash_{S5} (\Box\Diamond\neg p \wedge \Box p) \supset \Box\bot$ because $\vdash_{S5} \Box p \supset \Box\Box p$ and S5 is normal; likewise $\vdash_{S5} (\Box\Diamond p \wedge \Box\neg p) \supset \Box\bot$; finally, neither $\vdash_{S5} p$ nor $\vdash_{S5} \neg p$. These features of S5 follow easily from the fact that it is a consistent normal extension of $K4G_1$, the smallest normal logic Σ including both 4 and G_1 ($\Diamond\Box\alpha \supset \Diamond\Box\alpha$). Since the inconsistent logic is certainly not r.e. quasi-conservative, we have this generalization of Proposition 34.2:

Corollary 34.4 No normal extension of KG_14 is r.e. quasi-conservative.

We can use Corollary 34.1 to show several familiar weakenings of S5 not to be r.e. quasi-conservative. G_1 corresponds to the condition that accessibility be convergent, in the sense that if x and y are both accessible from w, then some world z is accessible from both x and y. Informally, G_1 says that agents either cognize that they do not cognize α or cognize that they do not cognize $\neg\alpha$. Any normal logic satisfying E also satisfies G_1, so Corollary 34.4 implies in particular the failure of r.e. quasi-conservativeness for the logics K4E and KD4E. Those two logics are the natural analogues for belief of S5 as a logic for knowledge, since they retain positive and negative introspection while dropping truthfulness altogether (K4E) or weakening it to consistency (KD4E). Thus they are often used as logics of belief. But positive and negative introspection together violate the computational constraint in a normal logic even in the absence of truthfulness. Thus, in a generalized context, K4E or KD4E impose unacceptably strong computational constraints as logics of belief, just as S5 does as a logic of knowledge.

For more examples, consider the schemas

$$B(\alpha \supset \Box\Diamond\alpha) \text{ and } D_1(\Box(\Box\alpha \supset \beta) \vee \Box(\Box\beta \supset \alpha)).$$

B corresponds to the condition that accessibility be symmetric, D_1 to the condition that accessibility be connected, in the sense that if x and y are both accessible from w, then either x is accessible from y or y is accessible from x. Any normal logic satisfying B or D_1 also satisfies G_1, so KB4 and KD_14 are not r.e. quasi-conservative. *A fortiori*, the same holds if one requires the agent to be consistent or truthful by adding D or T respectively. Thus KD4E, KDG14, KTG_14 (= S4.2), KDD_14 and KTD_14 (= S4.3) are also not r.e. quasi-conservative. All these are sublogics of S5; we shall need to weaken S5 considerably to find an r.e. quasi-conservative logic.

Theorem 34.3 is also applicable to logics without positive introspection. We can use T rather than 4 to derive $(\Box\Diamond\neg p \wedge \Box p) \supset \Box\bot$, so:

Corollary 34.5 No normal extension of KTG_1 is r.e. quasi-conservative.

Again, consider Alt_n $(\vee\{\Box(\wedge\{p_j : j<i\} \supset p_i) : i \leq n\})$, e.g., Alt_2 is $\Box p_0 \vee \Box(p_0 \supset p_1) \vee \Box((p_0 \wedge p_1) \supset p_2)$. Alt_n corresponds to the condition that from each world at most n worlds be accessible; informally, the agent rules out all but n specific possibilities. Setting $\alpha_i = \wedge\{p_j : j<i\} \supset p_i$ and $\beta_i = \neg\alpha_i$ in Theorem 34.3 gives:

Corollary 34.6 For any n, no r.e. quasi-conservative prenormal modal logic contains Alt_n.

An epistemic logic which imposes an upper bound on how many possibilities the agent can countenance thereby excludes the agent from having some consistent r.e. theories about the black box.

Some R.e. Conservative Logics

Since every modal logic with an r.e. [quasi-] conservative extension is itself r.e. [quasi-] conservative, an efficient strategy is to seek very strong r.e. [quasi-] conservative logics, even if they are implausibly strong for most epistemic applications, because we can note that the weaker and perhaps more plausible logics which they extend will also be r.e. [quasi-] conservative.

A large class of r.e. conservative logics arises as follows. Let Σ be any epistemic logic. The agent might cognize each theorem of Σ. Moreover, an epistemic logic Σ^* may imply this, in that $\vdash_{\Sigma^*} \Box\alpha$ whenever $\vdash_\Sigma \alpha$. Σ and Σ^* may be distinct, even incompatible. For example, let Ver be the smallest normal modal logic containing $\Box\bot$. Interpreted epistemically, Ver implies that the agent is inconsistent; but Ver itself is consistent. An epistemic theory consisting just of Ver falsely but consistently self-attributes inconsistency, and an epistemic logic may report that the agent self-attributes inconsistency without itself attributing inconsistency to the agent. Thus Ver* may contain $\Box\Box\bot$ without $\Box\bot$. Similarly, let Triv be the smallest normal modal logic containing all theorems of the form $\alpha \equiv \Box\alpha$. Interpreted epistemically, Triv implies that the agent cognizes that his beliefs contain all and only truths; but Triv itself does not contain all and only truths (neither $\vdash_{\mathrm{Triv}} p$ nor $\vdash_{\mathrm{Triv}} \neg p$). Thus

Triv* may contain $\Box(p \equiv \Box p)$ without $p \equiv \Box p$. To be more precise, for any modal logics Λ and Σ let $\Lambda\Box\Sigma$ be the smallest normal extension of Λ containing $\{\Box\alpha : \vdash_\Sigma \alpha\}$. We will prove that if Σ is consistent and normal then $K\Box\Sigma$ is r.e. conservative. $K\Box\Sigma$ is an epistemic logic for theorizing about theories that incorporate the epistemic logic Σ. R.e. conservativeness implies no constraint on what epistemic logic the agent uses beyond consistency (if Σ is inconsistent, then $K\Box\Sigma$ contains Alt_0 and so is not even r.e. quasi-conservative). In particular, the smallest normal logic K itself is r.e. conservative. Moreover, if Σ is consistent and normal, then $K4\Box\Sigma$ is r.e. conservative; that is, we can add positive introspection. In particular, K4 itself is r.e. conservative. We prove this by proving that K□Ver and K□Triv are r.e. conservative. Since K□Ver and K□Triv contain $\Box\Box\bot$ and $\Box(p \equiv \Box p)$ respectively, they are too strong to be useful epistemic logics themselves, but equally they are strong enough to contain many other logics of epistemic interest, all of which must also be r.e. conservative. By contrast, Ver and Triv are not themselves even r.e. quasi-conservative, for $\vdash_{\mathrm{Ver}} \mathrm{Alt}_0$ and $\vdash_{\mathrm{Triv}} \mathrm{Alt}_1$.

For future reference, call a mapping ϕ from $L_\Box$ into $L_\Box$ *respectful* if and only if $\phi p = p$ for all propositional variables p, $\phi\bot = \bot$ and $\phi(\alpha \supset \beta) = \phi\alpha \supset \phi\beta$ for all formulas α and β.

Lemma 34.7 K□Triv is r.e. conservative.

Proof Let R be an r.e. theory in L. Let δ and κ be respectful mappings from $L_\Box$ to L such that $\delta\Box\alpha = \delta\alpha$; $\kappa\Box\alpha = \top$ if $R \vdash_{PC} \delta\alpha$ and $\kappa\Box\alpha = \bot$ otherwise for all formulas α. (i) Axiomatize Triv with all truth-functional tautologies and formulas of the form $\alpha \equiv \Box\alpha$ as the axioms and MP as the only rule of inference (schema K and rule RN are easily derivable). By an easy induction on the length of proofs, $\vdash_{\mathrm{Triv}} \alpha$ only if $\vdash_{PC} \delta\alpha$. (ii) Axiomatize K□Triv with all truth-functional tautologies and formulas of the forms $\Box(\alpha \supset \beta) \supset (\Box\alpha \supset \Box\beta)$ and $\Box\gamma$ whenever $\vdash_{\mathrm{Triv}} \gamma$ as the axioms and MP as the only rule of inference (RN is a derived rule; its conclusion is always an axiom because the logic so defined is a sublogic of Triv). We show by induction on the length of proofs that $\vdash_{K\Box\mathrm{Triv}} \alpha$ only if $\vdash_{PC} \kappa\alpha$. Basis: If $\vdash_{PC} \alpha$, $\vdash_{PC} \kappa\alpha$. If $\kappa\Box(\alpha \supset \beta) = \top$ and $\kappa\Box\alpha = \top$ then $R \vdash_{PC} \delta\alpha \supset \delta\beta$ and $R \vdash_{PC} \delta\alpha$, so $R \vdash_{PC} \delta\beta$, so $\kappa\Box\beta = \top$, so $R \vdash_{PC} \kappa(\Box(\alpha \supset \beta) \supset (\Box\alpha \supset \Box\beta))$; otherwise $\kappa\Box(\alpha \supset \beta) = \bot$ or $\kappa\Box\alpha = \bot$ and again $R \vdash_{PC} \kappa(\Box(\alpha \supset \beta) \supset (\Box\alpha \supset \Box\beta))$. If $\vdash_{\mathrm{Triv}} \gamma$ then $\vdash_{PC} \delta\gamma$ by (i), so $R \vdash_{PC} \delta\gamma$, so $\kappa\Box\gamma = \top$, so $\vdash_{PC} \kappa\Box\gamma$. Induction step: trivial. (iii) Put $Y = \{\Box\alpha \in L_\Box : R \vdash_{PC} \delta\gamma\} \cup \{\neg\Box\alpha \in L_\Box : \text{not } R \vdash_{PC} \delta\alpha\}$. Y is K□Triv-consistent, for if $Y_0 \subseteq Y$ is finite and $\vdash_{K\Box\mathrm{Triv}} \wedge Y_0 \supset \bot$ then $\vdash_{PC} \kappa(\wedge Y_0 \supset \top)$ by (ii), i.e. $\vdash_{PC} \wedge\{\kappa\alpha : \alpha \in Y_0\} \supset \bot$, which is impossible since $\{\kappa\alpha : \alpha \in Y\} \subseteq \{\top, \neg\bot\}$. Let X be a maximal K□Triv-consistent extension of Y. By definition of Y, $\Box^{-1}X = \{\alpha : R \vdash_{PC} \delta\alpha\}$, which is r.e. because R is r.e. and δ is recursive (although κ need not be). If $\alpha \in L$, $\delta\alpha = \alpha$, so $\Box\alpha \in X$ if and only if $R \vdash_{PC} \alpha$, i.e., if and only if $\alpha \in R$ because R is a theory; thus $L \cap \Box^{-1}X = R$. Hence K□Triv is r.e. conservative.

Lemma 34.8 K□Ver is r.e. conservative.

Proof Like Lemma 34.7, but in place of δ use a respectful mapping λ such that $\lambda\Box\alpha = \top$.

A notable sublogic of K□Ver is GL, the smallest normal modal logic including $\Box(\Box\alpha \supset \alpha) \supset \Box\alpha$. Thus a corollary of Lemma 34.8 is that GL is r.e. conservative. GL is in a precise sense the logic of what is provable in Peano arithmetic (PA) about provability in PA (Boolos 1993 has exposition and references). More generally, if R is an ω-consistent r.e. extension of PA, then GL is the logic of what is provable in R about provability in R. Since a Turing machine's theory of arithmetic is presumably at best an ω-consistent r.e. extension of PA, GL is therefore a salient epistemic logic for Turing machines, and its r.e. conservativeness is not surprising.

Caution We must be careful in our informal renderings of results about provability logic. A provability operator creates an intensional context within which the substitution of coextensive but not provably coextensive descriptions can alter the truth-value of the whole sentence; this point applies in particular to descriptions of agents or their theories. On a provability interpretation of □, occurrences of □ within the scope of other occurrences of □ in effect involve just such occurrences of descriptions of agents or their theories in an intensional context, so which logic is validated can depend on the manner in which a given agent or theory is described. The validity of GL as an epistemic logic is relative to a special kind of descriptive self-presentation of the theory T in the interpretation of □, by a coding of its axioms and rules of inference. GL is not valid relative to some extensionally equivalent but intensionally distinct interpretations of □, e.g. the indexical reading 'I can prove that' as uttered by an epistemic subject with the computational capacity of a Turing machine (Shin and Williamson 1994; Williamson 1996, 1998).

Proposition 34.9 If Σ is a consistent normal modal logic, $K\Box\Sigma$ and $K4\Box\Sigma$ are r.e. conservative.

Proof By Makinson (1971), either $\Sigma \subseteq$ Triv or $\Sigma \subseteq$ Ver. Hence either $K\Box\Sigma \subseteq K\Box$Triv or $K\Box\Sigma \subseteq K\Box$Ver. But Schema 4 is easily derivable in both K□Triv and K□Ver, so $K4\Box\Sigma \subseteq K\Box$Triv or $K4\Box\Sigma \subseteq K\Box$Ver. By Lemmas 34.7 and 34.8, K□Triv and K□Ver are r.e. conservative, so $K4\Box\Sigma$ is.

All the logics salient in this paper are decidable, and therefore r.e., but we should note that an epistemic logic need not be r.e. to be r.e. conservative:

Corollary 34.10 Not all r.e. conservative normal modal logics are r.e.

Proof (i) We show that for any normal modal logic Σ, $\vdash_\Sigma \alpha$ if and only if $\vdash_{K\Box\Sigma} \Box\alpha$. Only the $\Leftarrow$ direction needs proving. Axiomatize $K\Box\Sigma$ with all truth-functional tautologies and formulas of the forms $\Box(\alpha \supset \beta) \supset (\Box\alpha \supset \Box\beta)$ and $\Box\gamma$ whenever $\vdash_\Sigma \gamma$ as the axioms and MP as the only rule of inference (RN is a derived rule; its conclusion is always an axiom because the logic so defined is a sublogic of Σ). Let η be a respectful mapping from $L_\Box$ to $L_\Box$ such that $\eta\Box\alpha = \alpha$ for all formulas α (η is distinct from δ in the proof of Lemma 34.7 since $\eta\Box\Box p = \Box p$ whereas $\delta\Box\Box p = p$). By induction on the length of proofs, $\vdash_{K\Box\Sigma} \alpha$ only if $\vdash_\Sigma \eta\alpha$. Hence $\vdash_{K\Box\Sigma} \Box\alpha$ only if $\vdash_\Sigma \alpha$. (ii) By (i), for any normal modal logics Σ_1 and Σ_2, $K\Box\Sigma_1 = K\Box\Sigma_2$ if and only if $\Sigma_1 = \Sigma_2$. But there are continuum many consistent normal modal logics (Blok 1980 has much more on these lines). Hence there are

continuum many corresponding logics of the form $K\Box\Sigma$; all are r.e. conservative by Proposition 34.9. Since only countably many modal logics are r.e., some of them are not r.e.

One limitation of Proposition 34.9 is that $K\Box\Sigma$ and $K4\Box\Sigma$ never contain the consistency schema D. In a sense this limitation is easily repaired. For any modal logic Σ, let $\Sigma\,[D]$ be the smallest extension of Σ containing D; thus $\vdash_{\Sigma[D]} \alpha$ just in case $\vdash_{\Sigma} \Diamond\top \supset \alpha$.

Proposition 34.11 For any r.e. conservative modal logic Σ, $\Sigma[D]$ is r.e. quasi-conservative.

Proof For any consistent theory R, any maximal Σ-consistent set X such that $L \cap \Box^{-1}X = R$ is $\Sigma[D]$-consistent because $\Diamond\top \in X$.

Corollary 34.12 If Σ is a consistent normal modal logic, $(K\Box\Sigma)[D]$ and $(K4\Box\Sigma)[D]$ are r.e. quasi-conservative.

Proof By Propositions 34.9 and 34.11.

Although $\Sigma[D]$ is always prenormal, it may not be normal, even if Σ is normal; sometimes not $\vdash_{\Sigma[D]} \Box\Diamond\top$. But we can also consider epistemic interpretations of normal logics with the D schema, e.g., KD and KD4. Such logics contain $\Box\Diamond\top$; they require agents to cognize their own consistency. By Gödel's second incompleteness theorem, this condition cannot be met relative to a Gödelian manner of representing the theory in itself; no consistent normal extension of the provability logic GL contains D. But $\Box\Diamond\top$ is true on other epistemic interpretations; for example, we know that our knowledge (as opposed to our beliefs) does not imply a contradiction. Since $GL \subseteq K\Box Ver$, Proposition 34.9 does not generalize to the r.e. quasi-conservativeness of $KD\Box\Sigma$. But we can generalize Lemma 34.7 thus:

Proposition 34.13 If $\Sigma \subseteq Triv$ then $KD\Box\Sigma$ and $K4\Box\Sigma$ are r.e. quasi-conservative.

Proof It suffices to prove that $KD\Box Triv$ (=$K4\Box Triv$) is r.e. quasi-conservative. Argue as for Lemma 34.1, adding $\Diamond\top$ as an axiom for $KD\Box Triv$ and noting that if R is consistent then $\kappa\Box\neg\top = \bot$, so $\vdash_{PC} \kappa\Diamond\top$.

In particular, KD and KD4 are themselves r.e. quasi-conservative; they are our first examples of r.e. quasi-conservative logics which are not r.e. conservative.

We now return to systems with the T schema. Since T implies D, only r.e. quasi-conservativeness is at issue. That constraint was motivated by the idea that any consistent r.e. theory in the non-modal language might be exactly the restriction of the agent's total r.e. theory to the non-modal language. On many epistemic interpretations, it is in the spirit of this idea that the agent's total theory might be true in the envisaged situation (for example, the agent's theory about the black box might be true, having been derived from a reliable witness). To require an epistemic logic Σ to leave open these possibilities is to require that $\Sigma[T]$ be r.e. quasi-conservative, where $\Sigma[T]$ is the smallest extension of Σ containing all instances of T. As with $\Sigma[D]$, $\Sigma[T]$ need not be normal even when Σ is; sometimes not

$\vdash_{\Sigma[T]} \Box(\Box\alpha \supset \alpha)$ (Williamson 1998, 113–116 discusses logics of the form Σ[T]). Agents may not cognize that they cognize only truths. Nevertheless, particularly when $\Box$ is interpreted in terms of knowledge, one might want an epistemic logic such as KT containing $\Box(\Box\alpha \supset \alpha)$.

Proposition 34.11 and Corollary 34.12 have no analogues for T in place of D. For any modal logic Σ, if $\vdash_\Sigma \alpha$ then $\vdash_{(K\Box\Sigma)[T]} \Box\alpha$, but $\vdash_{(K\Box\Sigma)[T]} \Box\alpha \supset \alpha$, so $\vdash_{(K\Box\Sigma)[T]} \alpha$; thus $(K\Box\Sigma)$[T] extends Σ and is r.e. quasi-conservative only if Σ is. Similarly, Proposition 34.12 would be false with T in place of D (counterexample: $\Sigma = S5$). Therefore, needing a different approach, we start with the system GL[T]. GL[T] has intrinsic interest, for it is the provability logic GLS introduced by Solovay and shown by him to be the logic of what is true (rather than provable) about provability in PA; more generally, it is the logic of what is true about provability in an ω-consistent r.e. extension of PA. GLS is therefore a salient epistemic logic for Turing machines, and its r.e. quasi-conservativeness is not surprising. Although GLS is not normal and has no consistent normal extension, we can use its r.e. quasi-conservativeness to establish that of normal logics containing T.

Proposition 34.14 GLS is r.e. quasi-conservative.

Proof Let R be an consistent r.e. theory in L. Axiomatize a theory R+ in $L_\Box$ with all members of R, truth-functional tautologies and formulas of the forms $\Box(\alpha \supset \beta) \supset (\Box\alpha \supset \Box\beta)$ and $\Box(\Box\alpha \supset \alpha) \supset \Box\alpha$ as the axioms and MP and RN as the rules of inference. Since R is r.e., so is R+. Let λ be the respectful mapping such that $\lambda\Box\alpha = \top$ for all formulas α. By an easy induction on the length of proofs, if $\vdash_{R+} \alpha$ then $R \vdash_{PC} \lambda\alpha$. But if $\alpha \in L$ then $\lambda\alpha = \alpha$, so $\vdash_{R+} \alpha$ only if $R \vdash_{PC} \alpha$, i.e., $\alpha \in R$; conversely, if $\alpha \in R$ then $\vdash_{R+} \alpha$; thus $L \cap R+ = R$. Let $Y \subseteq L$ be a maximal consistent extension of R. Define a set $X \subseteq L_\Box$ inductively: $p_i \in X \iff p_i \in Y$; $\bot \notin X$; $\alpha \supset \beta \in X \iff \alpha \notin X$ or $\beta \in X$; $\Box\alpha \in X \iff \vdash_{R+} \alpha$. For $\alpha \in L_\Box$, either $\alpha \in X$ or $\neg\alpha \in X$. We show by induction on the length of proofs that if $\vdash_{R+} \alpha$ then $\alpha \in X$. Basis: If $\alpha \in R$ then $\alpha \in Y \subseteq X$. If $\Box(\alpha \supset \beta) \in X$ and $\Box\alpha \in X$ then $\vdash_{R+} \alpha \supset \beta$ and $\vdash_{R+} \alpha$, so $\vdash_{R+} \beta$, so $\Box\beta \in X$; thus $\Box(\alpha \supset \beta) \supset (\Box\alpha \supset \Box\beta) \in X$. If $\Box(\Box\alpha \supset \alpha) \in X$ then $\vdash_{R+} \Box\alpha \supset \alpha$, so $\vdash_{R+} \Box(\Box\alpha \supset \alpha)$ because R+ is closed under RN; but $\vdash_{R+} \Box(\Box\alpha \supset \alpha) \supset \Box\alpha$, so $\vdash_{R+} \Box\alpha$, so $\vdash_{R+} \alpha$, so $\Box\alpha \in X$; thus $\Box(\Box\alpha \supset \alpha) \supset \Box\alpha \in X$. Induction step: Trivial. Now axiomatize GLS with all theorems of GL and formulas of the form $\Box\alpha \supset \alpha$ as the axioms and MP as the only rule of inference. We show by induction on the length of proofs that, for all formulas α, if $\vdash_{GLS} \alpha$ then $\alpha \in X$. Basis: If $\vdash_{GL} \alpha$ then $\vdash_{R+} \alpha$ because $GL \subseteq R+$, so $\alpha \in X$ by the previous induction. If $\Box\alpha \in X$ then $\vdash_{R+} \alpha$, so again $\alpha \in X$; thus $\Box\alpha \supset \alpha \in X$. Induction step: Trivial. Hence $GLS \subseteq X$, so X is maximal GLS-consistent. Now $L \cap \Box^{-1}X = L \cap R+ = R$ and $\Box^{-1}X = R+$ is r.e. Thus GLS is r.e. quasi-conservative.

We can extend Proposition 34.14 to another system of interest in relation to provability logic. Grz is the smallest normal modal logic containing all formulas of the form $\Box(\Box(\alpha \supset \Box\alpha) \supset \alpha) \supset \alpha$. Grz turns out to be in a precise sense the logic of what is both provable and true in PA (Boolos 1993, 155–161 has all the facts about Grz used here). Grz is intimately related to GLS in a way which allows us to extend the r.e. quasi-conservativeness of GLS to Grz:

Proposition 34.15 Grz is r.e. quasi-conservative.

Proof Let R be a consistent theory in L. By Proposition 34.14, for some maximal GLS-consistent X, $L \cap \Box^{-1}X = R$ and $\Box^{-1}X$ is r.e. Let τ be the respectful mapping from $L_\Box$ to $L_\Box$ such that $\tau\Box\alpha = \Box\tau\alpha \wedge \tau\alpha$ for all formulas α. Put $\tau^{-1}X = \{\alpha : \tau\alpha \in X\}$. Now $Grz \subseteq \tau^{-1}X$, for $\vdash_{Grz} \alpha$ if and only if $\vdash_{GLS} \tau\alpha$ (Boolos 1993, 156), so $\tau\alpha \in X$ since X is maximal GLS-consistent, so $\alpha \in \tau^{-1}X$. Since X is maximal GLS-consistent, $\tau^{-1}X$ is maximal Grz-consistent. Suppose $\alpha \in L$, so $\tau\alpha = \alpha$, so $\tau\Box\alpha = \Box\alpha \wedge \alpha$; if $\Box\alpha \in X$ then $\alpha \in X$ because $\vdash_{GLS} \Box\alpha \supset \alpha$, so $\tau\Box\alpha \in X$, so $\Box\alpha \in \tau^{-1}X$; conversely, if $\Box\alpha \in \tau^{-1}X$ then $\tau\Box\alpha \in X$, so $\Box\alpha \in X$. Thus $L \cap \Box^{-1}\ \tau^{-1}X = L \cap \Box^{-1}\ X = R$. Moreover, $\Box^{-1}\ \tau^{-1}X$ is r.e. because X is r.e. and τ is recursive. Thus Grz is r.e. quasi-conservative.

Grz is not plausible as the logic of other epistemic applications. It is not a sublogic of S5 and $\vdash_{Grz} \neg\Box(\Box p \wedge \Diamond\neg p)$, which in effect forbids agents to cognize that they do not cognize whether p is true. Yet you can know that what you know neither entails that the coin came down heads nor entails that it did not. However, since Grz extends the epistemically more plausible S4, the smallest normal modal logic including both the T and 4 schemas, its r.e. quasi-conservativeness entails that of S4. Truthfulness and positive introspection are together consistent with r.e. quasi-conservativeness.

Corollary 34.16 [Compare Shin and Williamson 1994 Proposition 4.] S4 is r.e. quasi-conservative.

Since S4 is r.e. quasi-conservative while S5, its extension by E, is not, and K4 is r.e. conservative while K4E is not, one might be tempted to blame E for the failure to satisfy the constraints, and to suppose that no normal logics with E is r.e. quasi-conservative. That would be a mistake; the next two propositions show that E is harmless when not combined with 4.

Proposition 34.17 KDE is r.e. quasi-conservative.

Proof Let R be a consistent r.e. theory in L. Let μ and θ be respectful mappings from $L_\Box$ to L such that for all formulas α, $\theta\Box\alpha = \top$ if $\vdash_{PC} \theta\alpha$ and $\theta\Box\alpha = \bot$ otherwise; $\mu\Box\alpha = \top$ if $R \vdash_{PC} \theta\alpha$ and $\mu\Box\alpha = \bot$ otherwise. Axiomatize KDE with all truth-functional tautologies and formulas of the forms $\Box(\alpha \supset \beta) \supset (\Box\alpha \supset \Box\beta)$, $\neg\Box\bot$ and $\neg\Box\alpha \supset \Box\neg\Box\alpha$ as the axioms and MP and RN as the rules of inference. We show by induction on the length of proofs that for all formulas α, $\vdash_{KDE} \alpha$ only if $\vdash_{PC} \theta\alpha$ and $\vdash_{PC} \mu\alpha$. Basis: If $R \vdash_{PC} \theta\alpha$, then $\mu\Box\alpha = \top$, so $\mu(\neg\Box\alpha \supset \Box\neg\Box\alpha) = \neg\top \supset \mu\Box\neg\Box\alpha$; if not, then not $\vdash_{PC} \theta\alpha$, so $\theta\Box\alpha = \bot$, so $\theta\Box\alpha = \neg\bot$, so $R \vdash_{PC} \theta\ \neg\Box\alpha$, so $\mu\Box\neg\Box\alpha = \top$, so $\mu(\neg\Box\alpha \supset \Box\neg\Box\alpha) = \neg\bot \supset \top$; either way, $\vdash_{PC} \mu(\neg\Box\alpha \supset \Box\neg\Box\alpha)$. The rest of the induction is by now routine. The rest of the proof is like that of Lemma 34.7, with θ and μ in place of δ and κ respectively.

Corollary 34.18 KE is r.e. conservative.

Proof KE is r.e. quasi-conservative by Proposition 34.17. Since not $\vdash_{KE} \Diamond\top$, KE is r.e. conservative by Proposition 34.1.

Although both positive and negative introspection are individually consistent with r.e. [quasi-] conservativeness, their conjunction is not. Part of the explanation is this: without positive introspection, an r.e. but non-recursive theory R can count as satisfying negative introspection by falsely equating the agent's theory with a recursive subtheory of R; the idea behind the clause for $\theta\Box\alpha$ in the proof of Proposition 34.17 is to use PC as such a subtheory. That R satisfies negative introspection by making false equations is crucial, for KE[T] is S5 itself. Although both negative introspection and truthfulness are individually consistent with r.e. [quasi-] conservativeness, their conjunction is not.

Related Non-computational Constraints

Although r.e. conservativeness and r.e. quasi-conservativeness are defined in computational terms, something remains when the computational element is eliminated. For given any [consistent] theory R in L, r.e. or not, we might require an epistemic logic to leave open the possibility that R is exactly the restriction of the agent's theory to L. On this view, an epistemic logic should impose no constraint beyond consistency on the agent's non-epistemic theorizing. Thus we define a modal logic Σ to be *conservative* if and only if for every theory R in L, $L \cap \Box^{-1}X = R$ for some maximal Σ-consistent set X. Σ is *quasi-conservative* if and only if for every consistent theory R in L, $L \cap \Box^{-1}X = R$ for some maximal Σ-consistent set X. Equivalently, Σ is [quasi-] conservative if and only if for every [consistent] theory R in L, $\{\Box\alpha : \alpha \in R\} \cup \{\neg\Box\alpha : \alpha \in L - R\}$ is Σ-consistent. We can assess how far r.e. conservativeness and r.e. quasi-conservativeness are specifically computational constraints by comparing them with conservativeness and quasi-conservativeness respectively.

Theorem 34.19 A prenormal modal logic Σ is quasi-conservative if and only if for no $n \vdash_\Sigma \mathrm{Alt}_n$.

Proof ($\Rightarrow$) Suppose that $\vdash_\Sigma \mathrm{Alt}_n$. Put $X = \{\Box\alpha : \alpha \in PC\} \cup \{\neg\Box\alpha : \alpha \in L - PC\}$. For all $i \leq n$, not $\vdash_{PC} \wedge\{p_j : j < i\} \supset p_i$, so $\neg\Box(\wedge\{p_j : j < i\} \supset p_i) \in X$. Hence $X \vdash_\Sigma \neg\mathrm{Alt}_n$, so $X \vdash_\Sigma \bot$. Since PC is a theory in L, Σ is not quasi-conservative. ($\Leftarrow$) Suppose that R is a consistent theory in L and $\{\Box\alpha : \alpha \in R\} \cup \{\neg\Box\alpha : \alpha \in L - R\}$ is not $\Sigma -$ consistent. Thus for some $\alpha_0, \ldots, \alpha_m \in R$ and $\beta_0, \ldots, \beta_n \in L - R$ (such β_i exist because R is consistent), $\vdash_\Sigma \wedge\{\Box\alpha_i : i \leq m\} \supset \vee\{\Box\beta_i : i \leq n\}$. Let $i \leq n$; since $\alpha_0, \ldots, \alpha_m \in R$, $\beta_i \in L - R$ and R is a theory, it follows that for some valuation υ_i of L onto {0, 1} (where $\upsilon_i(\bot) = 0$ and $\upsilon_i(\gamma_1 \supset \gamma_2) = 1$ just in case $\upsilon_i(\gamma_1) \leq \upsilon_i(\gamma_2)$), $\upsilon_i(\alpha_j) = 1$ for all $j \leq m$ and $\upsilon_i(\beta_i) = 0$. Put $\upsilon_{n+1} = \upsilon_0$. Set $\delta_i = \wedge\{p_j : j < i\} \wedge \neg p_i$ for $i \leq n$ and $\delta_{n+1} = \wedge\{p_j : j \leq n\}$. Let σ be the substitution such that for all j, $\sigma\, p_j = \vee\{\delta_i : \upsilon_i(p_j) = 1, i \leq n + 1\}$. Since Σ is closed under US, $\vdash_\Sigma \wedge\{\Box\sigma\alpha_i : i \leq m\} \supset \vee\{\Box\sigma\beta_i : i \leq n\}$. We can prove by induction on the complexity of γ that for all $\gamma \in L$ and $i \leq n + 1$, if $\upsilon_i(\gamma) = 1$ then $\vdash_{PC} \delta_i \supset \sigma\gamma$ and if $\upsilon_i(\gamma) = 0$ then $\vdash_{PC} \delta_i \supset \neg\sigma\gamma$. Basis: Immediate by definition of σ, for $\vdash_{PC} \delta_i \supset \neg\delta_k$ whenever $i \neq k$. Induction step:

Routine. Now for $i \leq n+1$ and $j \leq m$, $\upsilon_i(\alpha_j) = 1$, so $\vdash_{PC} \delta_i \supset \sigma\alpha_j$; since $\vdash_{PC} \vee\{\delta_i : i \leq n+1\}$, $\vdash_{PC} \sigma\alpha_j$, so $\vdash_{PC} \top \supset \sigma\alpha_j$. Hence by prenormality $\vdash_\Sigma \Box\top \supset \Box\sigma\alpha_j$ and so $\vdash_\Sigma \Box\sigma\alpha_j$. Thus $\vdash_\Sigma \vee\{\Box\sigma\beta_i : i \leq n\}$. Moreover, for each $i \leq n$, $\upsilon_i(\beta_i) = 0$, so $\vdash_{PC} \delta_i \supset \neg\sigma\beta_i$, so $\vdash_{PC} \sigma\beta_i \supset (\wedge\{p_j : j < i\} \supset p_i)$, so $\vdash_\Sigma \Box\sigma\beta_i \supset \Box\ (\wedge\{p_j : j < i\} \supset p_i)$. Thus $\vdash_\Sigma \mathrm{Alt}_n$.

Proposition 34.20 A prenormal modal logic Σ is conservative if and only if Σ is quasi-conservative and not $\vdash_\Sigma \Diamond\top$.

Proof Like Proposition 34.1 with 'r.e.' omitted.

Thus S5 is a quasi-conservative normal modal logic which is not r.e. quasi-conservative; K4E is a conservative normal modal logic which is not r.e. conservative. Most of the examples given above of logics which are not r.e. [quasi-] conservative are [quasi-] conservative. It is the distinctively computational requirements of r.e. quasi-conservativeness and r.e. conservativeness which those logics fail to meet.

Corollary 34.21 Every r.e. quasi-conservative prenormal modal logic is quasi-conservative; every r.e. conservative prenormal modal logic is conservative.

Proof From Proposition 34.1, Corollary 34.6, Theorem 34.19 and Proposition 34.20.

Although quasi-conservativeness exceeds r.e. quasi-conservativeness in requiring an epistemic logic to leave open the possibility that the restriction of the subject's theory to the language L is any given non-r.e. theory in L, this requirement is met by any epistemic logic which leaves open the corresponding possibility for every consistent r.e. theory in L.

Conclusion

Our investigation has uncovered part of a complex picture. The line between those modal logics weak enough to be r.e. conservative or r.e. quasi-conservative and those that are too strong appears not to coincide with any more familiar distinction between classes of modal logics, although a solution to the problem left open in the section "Some non-r.e. quasi-conservative logics" about the converse of Theorem 34.3 might bring clarification. What we have seen is that some decidable modal logics in general use as logics of knowledge (such as S5) or belief (such as KD45 and K45) when applied in generalized settings impose constraints on epistemic agents that require them to exceed every Turing machine in computational power. For many interpretations of epistemic logic, such a constraint is unacceptably strong.

The problem is not the same as the issue of logical omniscience, since many epistemic logics (such as S4 and various provability logics) do not impose the unacceptably strong constraints, although they do impose logical omniscience. Interpretations that finesse logical omniscience by building it into the definition

of the propositional attitude that interprets the symbol $\Box$ do not thereby finesse the computational issue that we have been investigating. Nevertheless, the two questions are related, because the deductive closure of a recursively axiomatised theory is what makes its theorems computationally hard to survey. In particular, it can be computationally hard to check for *non*-theoremhood, which is what negative introspection and similar axioms require. In fact, negative introspection by itself turned out not to impose unacceptable computational requirements (Corollary 34.18), but its combination with independently more plausible axioms does so. Perhaps the issues raised in this paper will provide a more fruitful context in which to discuss some of the questions raised by the debate on logical omniscience and bounded rationality.

The results proved in the paper also suggest that more consideration should be given to the epistemic use of weaker modal logics that are r.e. conservative or quasi-conservative. The plausibility of correspondingly weaker axioms must be evaluated under suitable epistemic interpretations. Weaker epistemic logics present a more complex picture of the knowing subject, but also a more nuanced one, because they make distinctions that stronger logics erase. We have seen that the more nuanced picture is needed to express the limits in general cognition of creatures whose powers do not exceed those of every Turing machine.

Acknowledgements Material based on this paper was presented to colloquia of the British Society for the Philosophy of Science and the Computer Science Laboratory at Oxford. I thank participants in both for useful comments.

References

Blok, W. J. (1980). The lattice of modal logics: An algebraic investigation. *Journal of Symbolic Logic, 45*, 221–236.

Boolos, G. (1993). *The logic of provability*. Cambridge: Cambridge University Press.

Craig, W. (1953). On axiomatizability within a system. *Journal of Symbolic Logic, 18*, 30–32.

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. (1995). *Reasoning about knowledge*. Cambridge, MA: MIT Press.

Makinson, D. C. (1971). Some embedding theorems in modal logic. *Notre Dame Journal of Formal Logic, 12*, 252–254.

Shin, H. S., & Williamson, T. (1994). Representing the knowledge of Turing machines. *Theory and Decision, 37*, 125–146.

Skyrms, B. (1978). An immaculate conception of modality. *The Journal of Philosophy, 75*, 368–387.

Williamson, T. (1996). Self-knowledge and embedded operators. *Analysis, 56*, 202–209.

Williamson, T. (1998). Iterated attitudes. In T. Smiley (Ed.), *Philosophical logic* (pp. 85–133). Oxford: Oxford University Press for the British Academy.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

Part V
Interactive Epistemology

Chapter 35
Introduction

Horacio Arló-Costa, Vincent F. Hendricks, and Johan van Benthem

The last few decades have witnessed the growing importance of multi-agent perspectives in epistemic matters. While traditional epistemology has largely centered on what single agents know, barring the occasional encounter with a skeptic, there has been a growing focus on interaction in many disciplines, turning from single-reasoner to many-reasoners problems, the way physicists turned to many-body constellations as the essence of nature. This trend may be seen in social epistemology, speaker-hearer views of meaning, dialogical foundations of logic, or multi-agent systems instead of single computing devices in computer science. While an inference or an observation may be the basic informational act for a single agent, think of a question plus answer as the unit of social communication. This agent exchange involves knowledge about facts and about others, and the information that flows and thus changes the current epistemic state of both agents in systematic ways. Existing epistemic and doxastic logics can describe part of this setting, since they allow for iteration for different agents, expressing thinks like "agent 1 believes that

Horacio Arló-Costa was deceased at the time of publication.

H. Arló-Costa (deceased)
Carnegie Mellon University, Pittsburgh, PA, USA

V.F. Hendricks (✉)
Center for Information and Bubble Studies, University of Copenhagen, Copenhagen, Denmark
e-mail: vincent@hum.ku.dk

J. van Benthem
University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Stanford University, Stanford, United States
e-mail: johan@science.uva.nl

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_35

agent 2 knows whether the heater is on". But the next level of social interaction involves the formation of groups with their own forms of knowledge, based perhaps on shared information.

One crucial notion concerning groups is 'common knowledge', which has come up in several disciplines independently. The article by Lewis gives an early philosophical motivation and development for common knowledge as a foundation for coordinated behavior, while the piece by Barwise points out logical subtleties in defining this notion that still have not all been disentangled decades later. Another crucial phenomenon, as we said, is the nature of the dynamic actions and events that drive the flow of information and interaction. Such actions have been brought inside the scope of logic by combining ideas from epistemic logic and logics of programs in the seminal paper by Baltag, Solecki and Moss. Baltag and Smets generalize this social dynamics to include belief revision and related phenomena. In particular, these papers show how a wide variety of actions of knowledge update and belief change can be described in 'dynamic-epistemic logics' allowing for groups with individual differences in observational power and habits of revision.

All these things and more come together in the compass of games, where information-driven agents pursue goals based on their evaluation of outcomes, through an interaction over time involving individual strategies responding to what others do. Indeed, the foundations of game theory have been a 'philosophical lab' since the mid 1970s, and the classic paper by Aumann on agreeing to disagree shows how group knowledge, investigative procedure, and eventually communication of disagreement can be subject to surprising logical laws. Perhaps the most well-studied notions in game theory are epistemic conditions such as common knowledge or common belief in rationality guaranteeing, under broad conditions, that players follow the backward induction solution, or other important solution concepts in games. The paper by Aumann & Brandenburger provides key examples of this style of analysis, as a major representative of what is by now a large body of literature. The philosophical community has started picking up on these developments, and related ones in the study of multi-agent systems: partly due to their intrinsic interest, and partly because games form a concrete microcosm where about all issues that have occupied philosophical logicians occur together. The article by Stalnaker shows how this encounter of logic and game theory can be of clear mutual benefit, and the perceptive paper by Halpern analyzes the ensuing debate across communities about the status of rationality.

Suggested Further Reading

Epistemic foundations of game theory as initiated by Aumann is a very rich area. Here is a selection of just a few papers that will set the reader thinking: J. Geanakoplos & H. Polemarchakis, 'We Can't Disagree Forever', Journal of Economic Theory 28, 1982, 192–200; P. Battigalli & G. Bonanno, 'Recent Results on Belief, Knowledge and the Epistemic Foundations of Game Theory', Research in Economics 53(2), 1999, 149–225; A. Brandenburger, 'The Power of Paradox: some Recent Developments in Interactive Epistemology', International Journal of Game Theory 35(4): 465–492. D. Samet & P. Jehiel, 'Learning to Play Games in Extensive Form by Valuation', Journal of Economic Theory 124, 2005, 129–148.

Epistemic considerations on agency in computer science are an equally rich area. We have given some textbook references in connection with logics of knowledge and belief, but here is an influential classical paper: J. Y. Halpern & Y. Moses, 'Knowledge and Common Knowledge in a Distributed Environment', Journal of the ACM 37:3, 1990, 549–587. An excellent monograph tying together logics of agency with game theory and much more is Y. Shoham & K. Leyton-Brown, Multiagent Systems, Cambridge University Press, 2009. Zooming in on 'dynamic-epistemic' foundations of information-driven agency, a comprehensive treatment is in J. van Benthem, Logical Dynamics of Information and Interaction, Cambridge University Press, 2011. Additional specifics of the interface of logic and games are found in W. van der Hoek & M. Pauly, 'Modal Logic and Games', in P. Blackburn, J. van Benthem & F. Wolter, eds. Handbook of Modal Logic, Elsevier, 2006, 1077–1148. A comprehensive philosophical analysis is found in B. de Bruin, Explaining Games: The Epistemic Programme in Game Theory, Springer, 2010. Finally, there is the growing stream of social epistemology, an area with many philosophical motivations beyond game theory, with books such as A. Goldman, Knowledge in a Social World, Oxford University Press, 1999, and Ch. List & Ph. Petit, Group Agency, Oxford University Press, 2011. Another important stream has been the influence of evolutionary game theory as a paradigm for philosophical studies, with B. Skyrms' The Stag Hunt and the Evolution of Social Structure, Cambridge University Press, 2004, as an outstanding example of the insights emerging in this way.

Chapter 36
Convention (An Excerpt on Coordination and Higher-Order Expectations)

David Lewis

Sample Coordination Problems

Use of language belongs to a class of situations with a conspicuous common character: situations I shall call *coordination problems*. I postpone a definition until we have seen a few examples. We begin with situations that might arise between two people—call them "you" and "I."

(1) Suppose you and I both want to meet each other. We will meet if and only if we go to the same place. It matters little to either of us where (within limits) he goes if he meets the other there; and it matters little to either of us where he goes if he fails to meet the other there. We must each choose where to go. The best place for me to go is the place where you will go, so I try to figure out where you will go and to go there myself. You do the same. Each chooses according to his expectation of the other's choice. If either succeeds, so does the other; the outcome is one we both desired.

(4) Suppose several of us are driving on the same winding two-lane roads. It matters little to anyone whether he drives in the left or the right lane, provided the others do likewise. But if some drive in the left lane and some in the right, everyone is in danger of collision. So each must choose whether to drive in the left lane or in the right, according to his expectations about the others: to drive in the left lane if most or all of the others do, to drive in the right lane if most or all of the others do (and to drive where he pleases if the others are more or less equally divided).

This chapter is an excerpt of David Lewis' PhD thesis *Convention*. It contains the primary content of Chapter 1: *Coordination and Convention*, where Lewis presents the idea of conventions as a mutually expected regularity in coordination in a recurrent situation.

D. Lewis (deceased)
Princeton University, Princeton, NJ, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_36

(7) Suppose we are contented oligopolists. As the price of our raw material varies, we must each set new prices. It is to no one's advantage to set his prices higher than the others set theirs, since if he does he tends to lose his share of the market. Nor is it to anyone's advantage to set his prices lower than the others set theirs, since if he does he menaces his competitors and incurs their retaliation. So each must set his prices within the range of prices he expects the others to set.

Analysis of Coordination Problems

With these examples, let us see how to describe the common character of coordination problems.

Two or more agents must each choose one of several alternative actions. Often all the agents have the same set of alternative actions, but that is not necessary. The outcomes the agents want to produce or prevent are determined jointly by the actions of all the agents. So the outcome of any action an agent might choose depends on the actions of the other agents. That is why—as we have seen in every example—each must choose what to do according to his expectations about what the others will do.

To exclude trivial cases, a coordination problem must have more than one coordination equilibrium. But that requirement is not quite strong enough. Figure 36.1 shows two matrices in which, sure enough, there are multiple coordination equilibria (two on the left, four on the right). Yet there is still no need for either agent to base his choice on his expectation about the other's choice. There is no need for them to try for the same equilibrium—no need for coordination—since if they try for different equilibria, some equilibrium will nevertheless be reached. These cases exhibit another kind of triviality, akin to the triviality of a case with a unique coordination equilibrium.

A combination is an equilibrium if each agent likes it *at least as well as* any other combination he could have reached, given the others' choices. Let us call it a *proper* equilibrium if each agent likes it *better than* any other combination he could have

	C1	C2
R1	1 1	1 1
R2	0 0	0 0

	C1	C2	C3
R1	1 1	1 1	.2 0
R2	1 1	1 1	.5 .2
R3	0 .5	0 0	0 0

Fig. 36.1

	C1	C2	C3
R1	2 2	0 0	0 0
R2	0 0	2 2	0 0
R3	0 0	1 1	1 1

Fig. 36.2

reached, given the others' choices. In a two-person matrix, for instance, a proper equilibrium is preferred by Row-chooser to all other combinations in its column, and by Column-chooser to all other combinations in its row. In the matrices in Fig. 36.1, there are multiple coordination equilibria, but all of them are improper.

There is no need to stipulate that all equilibria in a coordination problem must be proper; it seems that the matrix in Fig. 36.2 ought to be counted as essentially similar to our clear examples of coordination problems, despite the impropriety of its equilibrium ⟨*R3, C3*⟩. The two proper coordination equilibria—⟨*R1*, *C1*⟩ and ⟨*R2, C2*⟩—are sufficient to keep the problem nontrivial. I stipulate instead that a coordination problem must contain at least two proper coordination equilibria.

This is only one—the strongest—of several defensible restrictions. We might prefer a weaker restriction that would not rule out matrices like those in Fig. 36.3. But a satisfactory restriction would be complicated and would entail too many qualifications later. And situations like those of Fig. 36.3 can be rescued even under the strong restriction we have adopted. Let *R2′* be the disjunction of *R2* and *R3*, and *C2′* the disjunction of *C2* and *C3* in the left-hand matrix. Then the same situation can be represented by the new matrix in Fig. 36.4, which does have two proper coordination equilibria. The right-hand matrix can be consolidated in a similar way. But matrices like the one in Fig. 36.5, which are ruled out by the strong restriction, and ought to be ruled out, cannot be rescued by any such consolidation.

To sum up: Coordination problems—situations that resemble my 11 examples in the important respects in which they resemble one another[1]—are situations of interdependent decision by two or more agents in which coincidence of interest predominates and in which there are two or more proper coordination equilibria. We could also say—though less informatively than one might think—that they are situations in which, relative to *some* classification of actions, the agents have a common interest in all doing the same one of several alternative actions.

[1]See Michael Slote, "The Theory of Important Criteria," *Journal of Philosophy*, 63 (1966), pp. 211–224. Slote shows that we commonly introduce a class by means of examples and take the defining features of the class to be those distinctive features of our examples which seem important for an understanding of their character. That is what I take myself to be doing here and elsewhere.

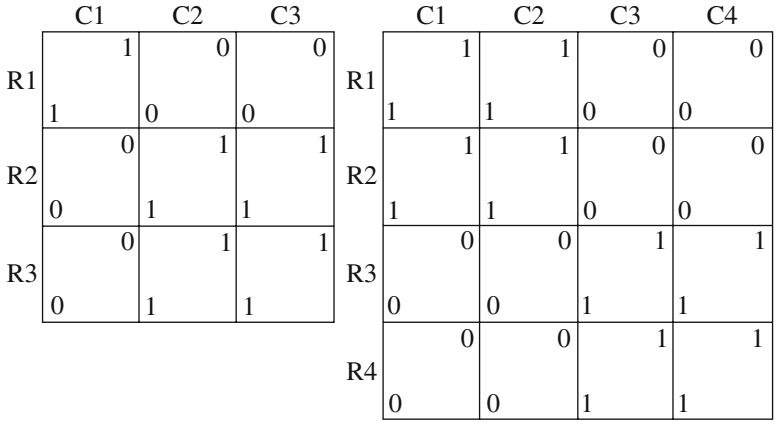


Fig. 36.3

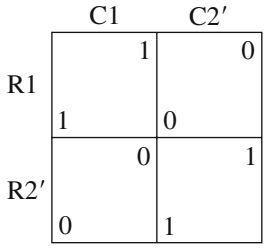


Fig. 36.4

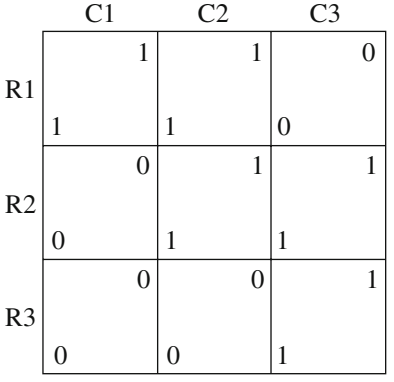


Fig. 36.5

Solving Coordination Problems

Agents confronted by a coordination problem may or may not succeed in each acting so that they reach one of the possible coordination equilibria. They might succeed just by luck, although some of them choose without regard to the others' expected actions (doing so perhaps because they cannot guess what the others will do, perhaps because the chance of coordination seems so small as to be negligible).

But they are more likely to succeed—if they do—through the agency of a system of suitably concordant mutual expectations. Thus in example (1) I may go to a certain place because I expect you to go there, while you go there because you expect me to; in example (2) I may call back because I expect you not to, while you do not because you expect me to; in example (4) each of us may drive on the right because he expects the rest to do so; and so on. In general, each may do his part of one of the possible coordination equilibria because he expects the others to do theirs, thereby reaching that equilibrium.

If an agent were completely confident in his expectation that the others would do their parts of a certain proper coordination equilibrium, he would have a decisive reason to do his own part. But if—as in any real case—his confidence is less than complete, he must balance his preference for doing his part if the others do theirs against his preferences for acting otherwise if they do not. He has a decisive reason to do his own part if he is *sufficiently* confident in his expectation that the others will do theirs. The degree of confidence which is sufficient depends on all his payoffs and sometimes on the comparative probabilities he assigns to the different *ways* the others might not all do their parts, in case not all of them do. For instance, in the coordination problem shown in Fig. 36.6, Row-chooser should do his part of the coordination equilibrium ⟨*R1*, *C1*⟩ by choosing *R1* if he has more than .5 confidence that Column-chooser will do his part by choosing *C1*. But in the coordination problems shown in Fig. 36.7, Row-chooser should choose R1 only if he has more than .9 confidence that Column-chooser will choose *C1*. If he has, say, .8 confidence that Column-chooser will choose *C1*, he would do better to choose *R2*, sacrificing his chance to achieve coordination at ⟨*R1*, *C1*⟩ in order to hedge against the possibility that his expectation was wrong. And in the coordination problem shown in Fig. 36.8, Row-chooser might be sure that if Column-chooser fails to do

	C1	C2
R1	1 1	0 0
R2	0 0	1 1

Fig. 36.6

	C1	C2
R1	1 1	−8 −8
R2	0 0	1 1

	C1	C2
R1	1 1	0 0
R2	0 0	9 9

	C1	C2
R1	3 3	−26 −26
R2	0 0	1 1

Fig. 36.7

	C1	C2	C3
R1	1 1	0 0	−8 −8
R2	0 0	1 1	9 9

Fig. 36.8

his part of ⟨*R1*, *C1*⟩, at least he will choose *C2*, not *C3*; if so, Row-chooser should choose *R1* if he has more than .5 confidence that Column-chooser will choose *C1*. Or Row-chooser might think that if Column-chooser fails to choose *R1*, he is just as likely to choose *C3* as to choose *C2*; if so, Row-chooser should choose *R1* only if he has more than .9 confidence that Column-chooser will choose *C1*. Or Row-chooser might be sure that if Column-chooser does not choose *C1*, he will choose *C3* instead; if so, Row-chooser's minimum sufficient degree of confidence is about .95. The strength of concordant expectation needed to produce coordination at a certain equilibrium is a measure of the difficulty of achieving coordination there, since however the concordant expectations are produced, weaker expectations will be produced more easily than stronger ones. (We can imagine cases in which so much mutual confidence is required to achieve coordination at an equilibrium that success is impossible. Imagine that a millionaire offers to distribute his fortune equally among a thousand men if each sends him $10; if even one does not, the millionaire will keep whatever he is sent. I take it that no matter what the thousand do to increase their mutual confidence, it is a practical certainty that the millionaire will not have to pay up. So if I am one of the thousand, I will keep my $10.)

We may achieve coordination by acting on our concordant expectations about each other's actions. And we may acquire those expectations, or correct or corroborate whatever expectations we already have, by putting ourselves in the other fellow's shoes, to the best of our ability. If I know what you believe about the matters of fact that determine the likely effects of your alternative actions, and if I know your

preferences among possible outcomes and I know that you possess a modicum of practical rationality, then I can replicate your practical reasoning to figure out what you will probably do, so that I can act appropriately.

In the case of a coordination problem, or any other problem of interdependent decision, one of the matters of fact that goes into determining the likely effects of your alternative actions is my own action. In order to figure out what you will do by replicating your practical reasoning, I need to figure out what *you* expect *me* to do.

I know that, just as I am trying to figure out what you will do by replicating your reasoning, so you may be trying to figure out what I will do by replicating my reasoning. This, like anything else you might do to figure out what I will do, is itself part of your reasoning. So to replicate your reasoning, I may have to replicate your attempt to replicate my reasoning.

This is not the end. I may reasonably expect *you* to realize that, unless I already know what you expect me to do, I may have to try to replicate your attempt to replicate my reasoning. So I may expect you to try to replicate my attempt to replicate your attempt to replicate my reasoning. So my own reasoning may have to include an attempt to replicate your attempt to replicate my attempt to replicate your attempt to replicate my reasoning. And so on.

Before things get out of hand, it will prove useful to introduce the concept of *higher-order expectations*, defined by recursion thus:

> A first-order expectation about something is an ordinary expectation about it.
> An $(n + 1)$th-order expectation about something $(n \geq 1)$ is an ordinary expectation about someone else's nth-order expectation about it.

For instance, if I expect you to expect that it will thunder, then I have a second-order expectation that it will thunder.

Whenever I replicate a piece of your practical reasoning, my second-order expectations about matters of fact, together with my first-order expectations about your preferences and your rationality, justify me in forming a first-order expectation about your action. In the case of problems of interdependent decision—for instance, coordination problems—some of the requisite second-order expectations must be about my own action.

Consider our first sample coordination problem: a situation in which you and I want to meet by going to the same place. Suppose that after deliberation I decide to come to a certain place. The fundamental practical reasoning which leads me to that choice is shown in Fig. 36.9. (In all diagrams of this kind, heavy arrows represent implications; light arrows represent causal connections between the mental states or actions of a rational agent.) And if my premise for this reasoning—my expectation that you will go there—was obtained by replicating your reasoning, my replication is shown in Fig. 36.10. And if my premise for this replication—my expectation that you will expect me to go there—was obtained by replicating your replication of my reasoning, my replication of your replication is shown in Fig. 36.11. And so on. The whole of my reasoning (simplified by disregarding the rationality premises) may be represented as in Fig. 36.12 for whatever finite number of stages it may take for

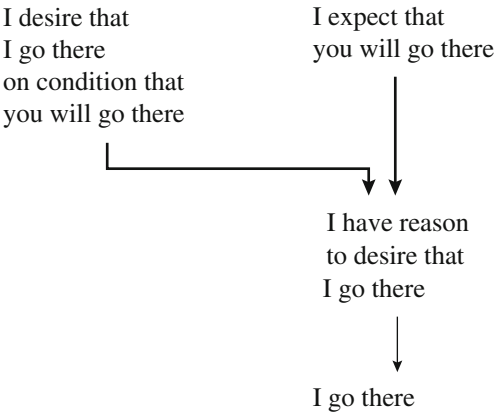


Fig. 36.9

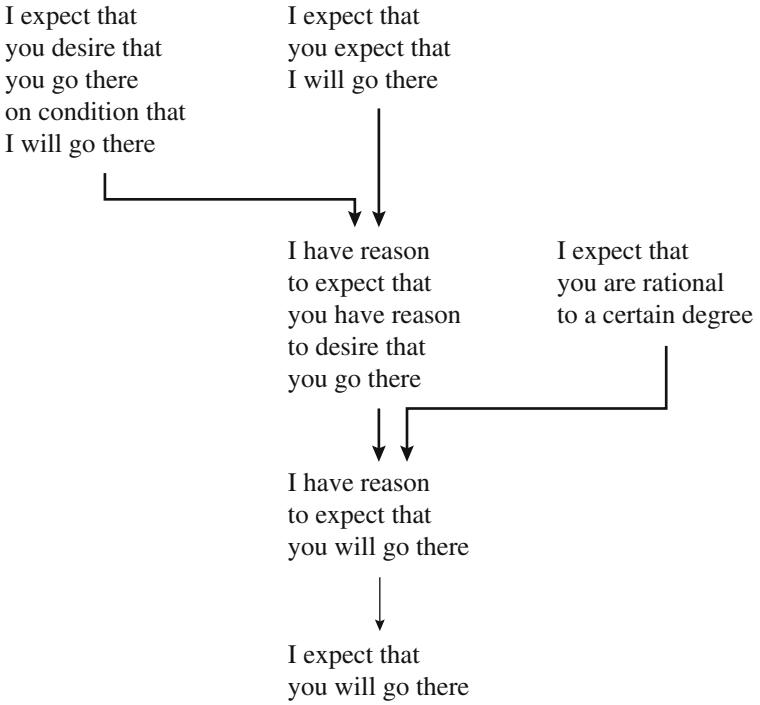


Fig. 36.10

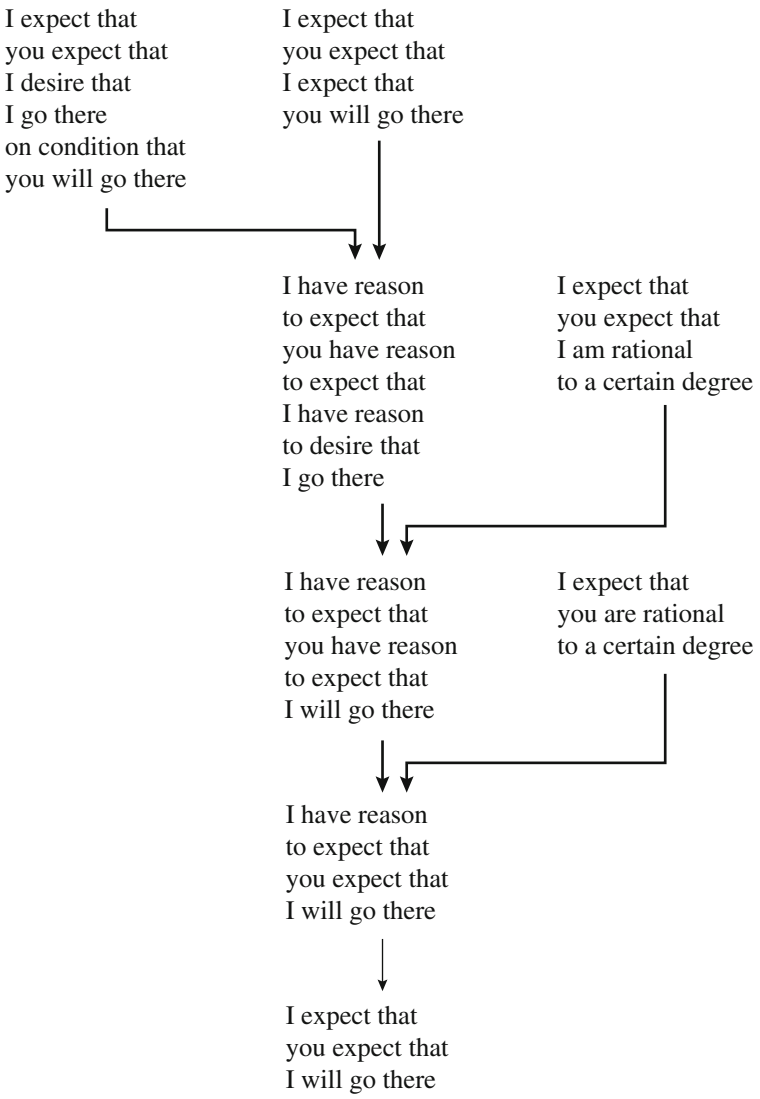


Fig. 36.11

me to use whatever higher-order expectations may be available to me regarding our actions and our conditional preferences. Replications are nested to some finite depth: my reasoning (outer boundary) contains a replication of yours (next boundary), which contains a replication of your replication of mine (next boundary), and so on.

So if I somehow happen to have an *n*th-order expectation about action in this two-person coordination problem, I may work outward through the nested replications to

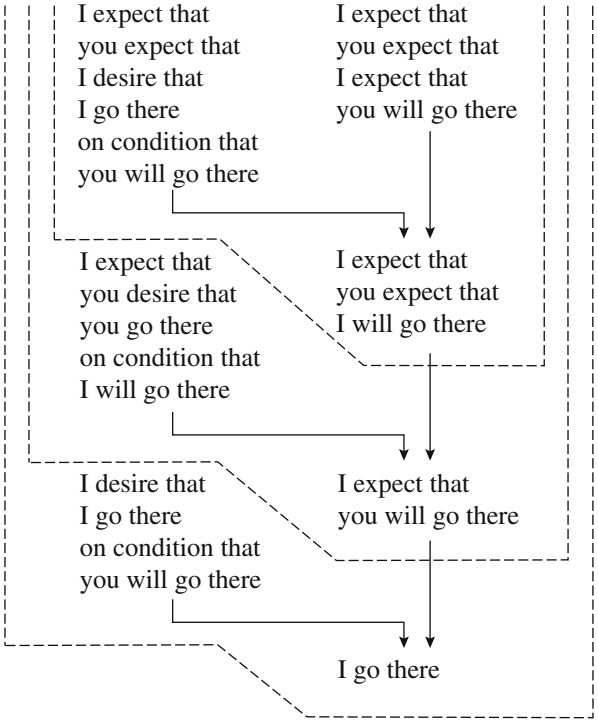


Fig. 36.12

lower- and lower-order expectations about action. Provided I go on long enough, and provided all the needed higher-order expectations about preferences and rationality are available, I eventually come out with a first-order expectation about your action—which is what I need in order to know how I should act.

Clearly a similar process of replication is possible in coordination problems among more than two agents. In general, my higher-order expectations about something are my expectations about x_1's expectations about x_2's expectations ... about it. (The sequence x_1, x_2 ... may repeat, but x_1 cannot be myself and no one can occur twice in immediate succession.) So when m agents are involved, I can have as many as $(m-1)^n$ different nth-order expectations about anything, corresponding to the $(m-1)^n$ different admissible sequences of length n. Replication in general is ramified: it is built from stages in which $m-1$ of my various $(n+1)$th-order expectations about action, plus ancillary premises, yield one of my nth-order expectations about action. I suppressed the ramification by setting $m=2$, but the general case is the same in principle.

Note that replication is *not* an interaction back and forth between people. It is a process in which *one* person works out the consequences of his beliefs about the world—a world he believes to include other people who are working out the

consequences of their beliefs, including their belief in other people who ... By our interaction in the world we acquire various high-order expectations that can serve us as premises. In our subsequent reasoning we are windowless monads doing our best to mirror each other, mirror each other mirroring each other, and so on.

Of course I do not imagine that anyone will solve a coordination problem by first acquiring a seventeenth-order expectation from somewhere and then sitting down to do his replications. For one thing, we rarely do have expectations of higher order than, say, fourth. For another thing, any ordinary situation that could justify a high-order expectation would also justify low-order expectations directly, without recourse to nested replications.

All the same, given the needed ancillary premises, an expectation of arbitrarily high order about action does give an agent *one* good reason for a choice of action. The one may, and normally will, be one reason among the many which jointly suffice to justify his choice. Suppose the agent is originally justified somehow in having expectations of several orders about his own and his partners' actions. And suppose the ancillary premises are available. Then each of his original expectations independently gives him a reason to act one way or another. If he is lucky, all these independent reasons will be reasons for the same action.[2] Then that action is strongly, because redundantly, justified; he has more reason to do it than could have been provided by any one of his original expectations by itself.

I said earlier that coordination might be rationally achieved with the aid of concordant mutual expectations about action. We have seen that these may be derived from first- and higher-order expectations about action, preferences, and rationality. So we generalize: coordination may be rationally achieved with the aid of a system of concordant mutual expectations, of first or higher orders, about the agents' actions, preferences, and rationality.

The more orders of expectation about action contribute to an agent's decision, the more independent justifications the agent will have; and insofar as he is aware of those justifications, the more firmly his choice will be determined. Circumstances that will help to solve a coordination problem, therefore, are circumstances in which the agents become justified in forming mutual expectations belonging to a concordant system. And the more orders, the better.

[2]Michael Scriven, in "An Essential Unpredictability in Human Behavior," *Scientific Psychology: Principles and Approaches,* ed. B. B. Wolman (New York: Basic Books, 1965), has discussed mutual replication of practical reasoning between agents in a game of conflict who want *not* to conform to each other's expectations. There is a cyclic alternation: from my (n + 4)th-order expectation that I will go to Minsk to my $(n + 3)$th-order expectation that you will go to Pinsk to my $(n + 2)$th-order expectation that I will go to Pinsk to my $(n + 1)$th-order expectation that you will go to Minsk to my nth-order expectation that I will go to Minsk ... Scriven notices that we cannot both act on complete and accurate replications of each other's reasoning. He takes this to prove human unpredictability. But perhaps it simply proves that the agents cannot both have enough time to finish their replications, since the time either needs increases with the time the other uses. See David Lewis and Jane Richardson, "Scriven on Human Unpredictability," *Philosophical Studies*, 17 (1966), pp. 69–74.

In considering how to solve coordination problems, I have postponed the answer that first comes to mind: by agreement. If the agents can communicate (without excessive cost), they can ensure a common understanding of their problem by discussing it. They can choose a coordination equilibrium—an arbitrary one, or one especially good for some or all of them, or one they can reach without too much mutual confidence. And each can assure the rest that he will do his part of the chosen equilibrium. Coordination by means of an agreement is not, of course, an alternative to coordination by means of concordant mutual expectations. Rather, agreement is one means of producing those expectations. It is an especially effective means, since it produces strong concordant expectations of several orders.

Suppose you and I want to meet tomorrow; today we happen to meet, and we make an appointment. Each thereby gives evidence of his interest in going where the other goes and of his intention to go to a certain place. By observing this evidence, we form concordant first-order expectations about each other's preferences and action. By observing each other observing it, we may also form concordant second-order expectations. By observing each other observing each other observing it, we may even form concordant third-order expectations. And so on; not forever, of course, but limited by the amount of reasoning we do and the amount we ascribe to each other—perhaps one or two steps more. The result is a system of concordant mutual expectations of several orders, conducive to coordination by means of replication.

The agents' agreement might be an exchange of formal or tacit promises. But it need not be. Even a man whose word is his bond can remove the promissory force by explicit disavowal, if not otherwise. An exchange of declarations of present intention will be good enough, even if each explicitly retains his right to change his plans later. No one need bind himself to act against his own interest. Rather, it will be in the interest of each to do just what he has led the others to expect him to do, since that action will be best for him if the others act on their expectations.

If one does consider himself bound by a promise, he has a second, independent incentive. His payoffs are modified, since he has attached the onus of promise breaking to all but one choice. Indeed, he may modify his payoffs so much by promising that the situation is no longer a coordination problem at all. For instance, the agent's promised action might become his dominant choice: he might wish to keep his promise no matter what, coordination or no coordination.

If such a strong promise is made publicly, the others will know that they must go along with the one who has promised, for they know what he will do. Such forceful promising is a way of getting rid of coordination problems, not a way of solving them.

Explicit agreement is an especially good and common means to coordination—so much so that we are tempted to speak of coordination otherwise produced as *tacit* agreement. But agreement (literally understood) is not the only source of concordant expectations to help us solve our coordination problems. We do without agreement by choice if we find ourselves already satisfied with the content and strength of our mutual expectations. We do without it by necessity if we have no way to

communicate, or if we can communicate only at a cost that outweighs our improved chance of coordination (say, if we are conspirators being shadowed).

Schelling has experimented with coordination problems in which the agents cannot communicate. His subjects know only that they share a common understanding of their problem—for instance, they may get instructions describing their problem and stating that everyone gets the same instructions. It turns out that sophisticated subjects in an experimental setting can often do very well—much better than chance—at solving novel coordination problems without communicating. They try for a coordination equilibrium that is somehow *salient:* one that stands out from the rest by its uniqueness in some conspicuous respect. It does not have to be uniquely *good*; indeed, it could be uniquely bad. It merely has to be unique in some way the subjects will notice, expect each other to notice, and so on. If different coordination equilibria are unique in different conspicuous ways, the subjects will need to be alike in the relative importance they attach to different respects of comparison; but often they are enough alike to solve the problem.

How can we explain coordination by salience? The subjects might all tend to pick the salient as a last resort, when they have no stronger ground for choice. Or they might expect each other to have that tendency, and act accordingly; or they might expect each other to expect each other to have that tendency and act accordingly, and act accordingly; and so on. Or—more likely—there might be a mixture of these. Their first- and higher-order expectations of a tendency to pick the salient as a last resort would be a system of concordant expectations capable of producing coordination at the salient equilibrium.

If their expectations did produce coordination, it would not matter whether anyone really would have picked the salient as a last resort. For each would have had a good reason for his choice, so his choice would not have been a last resort.

Thus even in a novel coordination problem—which is an extreme case—the agents can sometimes obtain the concordant expectations they need without communicating. An easier, and more common, case is that of a *familiar* coordination problem without communication. Here the agents' source of mutual expectations is precedent: acquaintance with past solved instances of their present coordination problem.

Convention

Let us start with the simplest case of coordination by precedent and generalize in various ways. In this way we shall meet the phenomenon I call *convention,* the subject of this book.

Suppose we have been given a coordination problem, and we have reached some fairly good coordination equilibrium. Given exactly the same problem again, perhaps each of us will repeat what he did before. If so, we will reach the same solution. If you and I met yesterday—by luck, by agreement, by salience, or however—and today we find we must meet again, we might both go back to

yesterday's meeting place, each hoping to find the other there. If we were cut off on the telephone and you happened to call back as I waited, then if we are cut off again in the same call, I will wait again.

We can explain the force of precedent just as we explained the force of salience. Indeed, precedent is merely the source of one important kind of salience: conspicuous uniqueness of an equilibrium because we reached it last time. We may tend to repeat the action that succeeded before if we have no strong reason to do otherwise. Whether or not any of us really has this tendency, we may somewhat expect each other to have it, or expect each other to expect each other to have it, and so on—that is, we may each have first- and higher-order expectations that the others will do their parts of the old coordination equilibrium, unless they have reason to act otherwise. Each one's expectation that the others will do their parts, strengthened perhaps by replication using his higher-order expectations, gives him some reason to do his own part. And if his original expectations of some order or other were strong enough, he will have a decisive reason to do his part. So he will do it.

I have been supposing that we are given a coordination problem, and then given the same problem again. But, of course, we could never be given exactly the same problem twice. There must be this difference at least: the second time, we can draw on our experience with the first. More generally, the two problems will differ in several independent respects. We cannot do exactly what we did before. Nothing we could do this time is exactly like what we did before—like it in every respect—because the situations are not exactly alike.

So suppose not that we are given the original problem again, but rather that we are given a new coordination problem analogous somehow to the original one. Guided by whatever analogy we notice, we tend to follow precedent by trying for a coordination equilibrium in the new problem which uniquely corresponds to the one we reached before.

There might be alternative analogies. If so, there is room for ambiguity about what would be following precedent and doing what we did before. Suppose that yesterday I called you on the telephone and I called back when we were cut off. Today you call me and we are cut off. We have a precedent in which I called back and a precedent—the same one—in which the original caller called back. But this time you are the original caller. No matter what I do this time, I do something analogous to what we did before. Our ambiguous precedent does not help us.

In fact, there are always innumerable alternative analogies. Were it not that we happen uniformly to notice some analogies and ignore others—those we call "natural" or "artificial," respectively—precedents would always be completely ambiguous and worthless. *Every* coordination equilibrium in our new problem (every other combination, too) corresponds uniquely to what we did before under *some* analogy, shares *some* distinctive description with it alone. Fortunately, most of the analogies are artificial. We ignore them; we do not tend to let them guide our choice, nor do we expect each other to have any such tendency, nor do we expect each other to expect each other to, and so on. And fortunately we have learned that all of us will mostly notice the same analogies. That is why precedents can be unambiguous in practice, and often are. If we notice only one of the analogies

between our problem and the precedent, or if one of those we notice seems far more conspicuous than the others, or even if several are conspicuous but they all happen to agree in indicating the same choice, then the other analogies do not matter. We are not in trouble unless conflicting analogies force themselves on our attention.

The more respects of similarity between the new problem and the precedent, the more likely it is that different analogies will turn out to agree, the less room there will be for ambiguity, and the easier it will be to follow precedent. A precedent in which I, the original caller, called back is ambiguous given a new problem in which you are the original caller—but not given a new problem in which I am again the original caller. That is why I began by pretending that the new problem was like the precedent in all respects.

Salience in general is uniqueness of a coordination equilibrium in a preeminently conspicuous respect. The salience due to precedent is no exception: it is uniqueness of a coordination equilibrium in virtue of its preeminently conspicuous analogy to what was done successfully before.

So far I have been supposing that the agents who set the precedent are the ones who follow it. This made sure that the agents given the second problem were acquainted with the circumstances and outcome of the first, and expected each other to be, expected each other to expect each other to be, and so on. But it is not an infallible way and not the only way. For instance, if yesterday I told you a story about people who got separated in the subway and happened to meet again at Charles Street, and today we get separated in the same way, we might independently decide to go and wait at Charles Street. It makes no difference whether the story I told you was true, or whether you thought it was, or whether I thought it was, or even whether I claimed it was. A fictive precedent would be as effective as an actual one in suggesting a course of action for us, and therefore as good a source of concordant mutual expectations enabling us to meet. So let us just stipulate that somehow the agents in the new problem are acquainted with the precedent, expect each other to be acquainted with it, and so on.

So far I have been supposing that we have a single precedent to follow. But we might have several. We might all be acquainted with a class of previous coordination problems, naturally analogous to our present problem and to each other, in which analogous coordination equilibria were reached. This is to say that the agents' actions conformed to some noticeable regularity. Since our present problem is suitably analogous to the precedents, we can reach a coordination equilibrium by all conforming to this same regularity. Each of us wants to conform to it if the others do; he has a *conditional preference* for conformity. If we do conform, the explanation has the familiar pattern: we tend to follow precedent, given no particular reason to do anything else; we expect that tendency in each other; we expect each other to expect it; and so on. We have our concordant first- and higher-order expectations, and they enable us to reach a coordination equilibrium.

It does not matter *why* coordination was achieved at analogous equilibria in the previous cases. Even if it had happened by luck, we could still follow the precedent set. One likely course of events would be this; the first case, or the first few, acted

as precedent for the next, those for the next, and so on. Similarly, no matter how our precedents came about, by following them this time we add this case to the stock of precedents available henceforth.

Several precedents are better than one, not only because we learn by repetition but also because differences between the precedents help to resolve ambiguity. Even if our present situation bears conflicting natural analogies to any one precedent, maybe only one of these analogies will hold between the precedents; so we will pay attention only to that one. Suppose we know of many cases in which a cut-off telephone call was restored, and in every case it was the original caller who called back. In some cases I was the original caller, in some you were, in some neither of us was. Now we are cut off and I was the original caller. For you to call back would be to do something analogous—under one analogy—to what succeeded in some of the previous cases. But we can ignore that analogy, for under it the precedents disagree.

Once there are many precedents available, without substantial disagreement or ambiguity, it is no longer necessary for all of us to be acquainted with precisely the same ones. It is enough if each of us is acquainted with some agreeing precedents, each expects everyone else to be acquainted with some that agree with his, each expects everyone else to expect everyone else to be acquainted with some precedents that agree with his, etc. It is easy to see how that might happen: if one has often encountered cases in which coordination was achieved in a certain problem by conforming to a certain regularity, and rarely or never encountered cases in which it was not, he is entitled to expect his neighbors to have had much the same experience. If I have driven all around the United States and seen many people driving on the right and never one on the left, I may reasonably infer that almost everyone in the United States drives on the right, and hence that this man driving toward me also has mostly seen people driving on the right—even if he and I have not seen any of the *same* people driving on the right.

Our acquaintance with a precedent need not be very detailed. It is enough to know that one has learned of many cases in which coordination was achieved in a certain problem by conforming to a certain regularity. There is no need to be able to specify the time and place, the agents involved, or any other particulars; no need to be able to recall the cases one by one. I cannot cite precedents one by one in which people drove on the right in the United States; I am not sure I can cite even one case; nonetheless, I know very well that I have often seen cars driven in the United States, and almost always they were on the right. And since I have no reason to think I encountered an abnormal sample, I infer that drivers in the United States do almost always drive on the right; so anyone I meet driving in the United States will believe this just as I do, will expect me to believe it, and so on.

Coordination by precedent, at its simplest, is this: achievement of coordination by means of shared acquaintance with the achievement of coordination in a single past case exactly like our present coordination problem. By removing inessential restrictions, we have come to this: achievement of coordination by means of shared acquaintance with a *regularity* governing the achievement of coordination in a class of past cases which bear some conspicuous analogy to one another and to our

present coordination problem. Our acquaintance with this regularity comes from our experience with some of its instances, not necessarily the same ones for everybody.

Given a regularity in past cases, we may reasonably extrapolate it into the (near) future. For we are entitled to expect that when agents acquainted with the past regularity are confronted by an analogous new coordination problem, they will succeed in achieving coordination by following precedent and continuing to conform to the same regularity. We come to expect conforming actions not only in past cases but in future ones as well. We acquire a general belief, unrestricted as to time, that members of a certain population conform to a certain regularity in a certain kind of recurring coordination problem for the sake of coordination.

Each new action in conformity to the regularity adds to our experience of general conformity. Our experience of general conformity in the past leads us, by force of precedent, to expect a like conformity in the future. And our expectation of future conformity is a reason to go on conforming, since to conform if others do is to achieve a coordination equilibrium and to satisfy one's own preferences. And so it goes—we're here because we're here because we're here because we're here. Once the process gets started, we have a metastable self-perpetuating system of preferences, expectations, and actions capable of persisting indefinitely. As long as uniform conformity is a coordination equilibrium, so that each wants to conform conditionally upon conformity by the others, conforming action produces expectation of conforming action and expectation of conforming action produces conforming action.

This is the phenomenon I call convention. Our first, rough, definition is:

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if, in any instance of S among members of P,
>
> (1) everyone conforms to R;
> (2) everyone expects everyone else to conform to R;
> (3) everyone prefers to conform to R on condition that the others do, since S is a coordination problem and uniform conformity to R is a proper coordination equilibrium in S.

Sample Conventions

Chapter II will be devoted to improving the definition. But before we hide the concept beneath its refinements, let us see how it applies to examples. Consider some conventions to solve our sample coordination problems.

(1) If you and I must meet every week, perhaps at first we will make a new appointment every time. But after we have met at the same time and place for a few weeks running, one of us will say, "See you here next week," at the end of every meeting. Later still we will not say anything (unless our usual arrangement is going to be unsatisfactory next week). We will just both go regularly to a certain place at a certain time every week, each going there to meet the other and confident that he will show up. This regularity that has gradually developed in our behavior is a convention.

Chapter 37
Three Views of Common Knowledge

Jon Barwise

Introduction

As the pioneering work of Dretske[1] has shown, knowing, believing, and having information are closely related and are profitably studied together. Thus while the title of this paper mentions common knowledge, I really have in mind the family of related notions including common knowledge, mutual belief and shared information. Even though I discuss common knowledge in this introduction, the discussion is really intended to apply to all three notions.

Common knowledge and its relatives have been written about from a wide variety of perspectives, including psychology, economics, game theory, computer science, the theory of convention, deterrence theory, the study of human-machine interaction, and the famous Conway paradox, just to mention a few. There are literally hundreds of papers that touch on the topic. However, while common knowledge is widely recognized to be an important phenomenon, there is no agreement as to just what it amounts to. Or rather, as we will see, what agreement there is presupposes a set of simplifying assumptions that are completely unrealistic. This paper offers a comparison of three competing views in a context which does not presuppose them to be equivalent, and explores their relationships in this context.[2]

Jon Barwise was deceased at the time of publication.

[1] F. Dretske, *Knowledge and the Flow of Information* (Cambridge, Mass.: Bradford Books/MIT Press, 1981).

[2] Obviously I have not read all, or even most, of the papers on common knowledge, so it could be that some or all of the points made in this paper are made elsewhere. If so, I would appreciate

J. Barwise (deceased)
Stanford University, Stanford, CA, USA

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_37

I take it that these accounts are after characterizations of common knowledge in terms of ordinary knowledge, of mutual belief in terms of belief, and of having shared information in terms of having information. Such accounts should be compatible with, but presumably distinct from, an account that shows how it is that common knowledge comes about. They should also be compatible with some explanation of how common knowledge is used.

We are going to compare the following approaches to common knowledge: (1) the *iterate* approach, (2) the *fixed-point* approach, and (3) the *shared-environment* approach. In order to review these three accounts, let's consider a special case where there are just two agents, say p and q, with common knowledge of some fact σ. Let τ be this additional fact, of the common knowledge of σ. We are looking for a characterization of τ in terms of p, q, σ and ordinary (private) knowledge.

By far the most common view of common knowledge is that τ is to be understood in terms of *iterated* knowledge of σ: p knows σ, q knows σ, p knows q knows σ, q knows p knows, p knows q knows p knows σ, and so forth. On this account, for p and q to have common knowledge of σ is for all members of this infinite collection of other facts to obtain. This is the approach taken in David Lewis' influential book[3] on convention, for example. It is, without doubt, the orthodox account, at least in the field of logic. It is, for example, the one that is the basis of the mathematical modeling of common knowledge in the logic of distributed systems.[4]

The two other accounts we want to investigate replace this infinite hierarchy with some sort of circularity. One such account was explicitly proposed by Harman.[5] Harman's proposal is that the correct analysis of τ is as:

$$p \text{ and } q \text{ know } (\sigma \text{ and } \tau)$$

Notice that on this fixed-point account, τ is in some sense a proper constituent of itself. Harman seems to suggest that this is nothing but a succinct representation of the first infinite hierarchy.

This fixed point approach is also the view of common knowledge that is implicit in Aumann's pioneering paper modeling common knowledge in game theory, as

learning about it. But even if this is so, I am reasonably sure that the particular model I develop below is original, depending as it does on recent work in set theory by Peter Aczel.

[3]David Lewis, *Convention*, *A Philosophical Study* (Cambridge, Mass.: Harvard University Press, 1969).

[4]See, for example, the paper by Halpern and Moses, "Knowledge and common knowledge in distributed environments," Proc. 3rd ACM Symp. on Principles of Distributed Computing (1984), 50–61, and the paper by Fagin, Halpern and Vardi, "A model-theoretic analysis of knowledge: preliminary report," Proc. 25th IEEE Symposium on Foundations of C.S., 268–278.

[5]See Gilbert Harman's review of *Linguistic Behavior* by Jonathan Bennett, *Language* 53 (1977): 417–24.

was pointed out by Tommy Tan and Sergio Ribeiro da Costa Werlang.[6] Aumann suggests that this approach is equivalent to the iterate approach. Tan and Ribeiro da Costa Werlang develop a mathematical model of the iterate approach and show that it is equivalent to Aumann's fixed point model. Similarly, one sees from the work of Halpern and Moses, that while they start with the iterate approach, in their set-up, this is equivalent to a fixed point. One of the aims of this paper is to develop a mathematical model where both iterate and fixed point accounts fit naturally, but where they are *not* equivalent. Only in such a framework can we explicitly isolate the assumptions that are needed to show them equivalent. We will see that these assumptions are simply false (in the case of knowledge), so that the issue as to which of the two, if either, is the "right" analysis of the notion is a live one.

The final approach we wish to discuss, the shared-environment approach, was proposed by Clark and Marshall,[7] in response to the enormous processing problems associated with the iterate account. On their account, p and q have common knowledge of σ just in case there is a situation s such that:

- $s \models \sigma$,
- $s \models p_1$ knows s,
- $s \models p_2$ knows s.

Here $s \models \theta$ is a notation for: θ is a fact of s. The intuitive idea is that common knowledge amounts to perception or other awareness of some situation, part of which includes the fact in question, but another part of which includes the very awarenesses of the situation by both agents. Again we note the circular nature of the characterization.

What Are We Modeling: Knowing or Having Information?

It is these three characterizations of common knowledge, and their relatives for the other notions of mutual belief and shared information, that we wish to compare. Among common knowledge, mutual belief, and shared information, we focus primarily on the case of having information, secondarily on the case of knowledge. Part of the claim of the paper is that these two notions are often conflated, and that it is this conflation that lends some credibility to the assumptions under which the first two approaches to common knowledge are equivalent. So I need to make clear

[6] R. J. Aumann, "Agreeing to disagree," *Annals of Statistics*, 4 (1976), 1236–1239, and the working paper "On Aumann's Notion of Common Knowledge – An alternative approach," Tan and Ribeiro da Costa Werlang. University of Chicago Graduate School of Business, 1986.

[7] H. Clark and C. Marshall, "Definite reference and mutual knowledge," in *Elements of Discourse Understanding*, ed. A. Joshi, B. Webber, and I. Sag (Cambridge: Cambridge University Press, 1981), 10–63.

what I take to be the difference between an agent p knowing some fact σ, and the agent simply having the information σ.

Here I am in agreement with Dretske.[8] Knowing σ is stronger than having the information σ. An agent knows σ if he not only has the information σ, but moreover, the information is "had" in a way that is tied up with the agent's abilities to act. When might this not be the case? The most notorious example (and by no means the only one) is when I know one fact σ, and another fact σ' logically follows from σ, but I disbelieve the latter because I don't know that the one follows from the other. Obviously there is a clear sense in which I have the information σ', but I certainly don't know it in the ordinary sense of the word. Another arises with certain forms of perceptual information. If I see the tallest spy hide a letter under a rock, then there is a clear sense in which I have the information the tallest spy has hidden the letter. However, if I don't know that he is a spy, say, then I don't know that the tallest spy has hidden a letter. Information travels at the speed of logic, genuine knowledge only travels at the speed of cognition and inference.

Much of the work in logic which seems to be about knowledge is best understood in terms of having information. And for good reason. For example, in dealing with computers, there is a good reason for our interest in the latter notion. We often use computers as information processors, after all, for our own ends. We are often interested less in what the computer does with the information it has, than in just what information it has and what we can do with it. Or, in the design of a robot, we may be aiming at getting the robot to behave in a way that is appropriate given the information it has. One might say, we are trying to make it know the information it has.

So, as noted earlier, this paper focuses primarily on the case of having information. The model I am going to develop originated with an analysis of shared perceptual information,[9] but it also works quite well for primary epistemic perceptional information[10] and the relation of having information.

Let me say all this in another way, since it seems to be a confusing point. In the section that follows, I could interpret the model as a model of knowledge if I were to make the same idealization that is made in most of the literature on common knowledge. However, part of what I want to do here is make very explicit just what the role of this idealization is in the modeling of common knowledge. Thus, I am forced to work in a context where we do not make it. Once we are clear about its role, we can then decide if we want to make it.

[8] Op. Cit.

[9] See ch. 2 of Fred Dretske, *Seeing and Knowing* (Chicago: University of Chicago Press, 1969); or J. Barwise, "Scenes and other Situations", *Journal of Philosophical Logic* 78 (1981): 369–97; or ch. 8 of J. Barwise and J. Perry, *Situations and Attitudes* (Cambridge, Mass.: Bradford Books/MIT Press, 1983).

[10] See ch. 3 of *Seeing and Knowing* or ch. 9 of *Situations and Attitudes.*

Summary of Results

Our results suggest that the fixed point approach gives the right theoretical analysis of the pretheortic notion of common knowledge. On the other hand, the shared-environment approach is the right way to understand how common knowledge usually arises and is maintained over an extended interaction. It does not offer an adequate characterization of the pretheoretic notion, though, since a given piece of common knowledge may arise from many different kinds of shared environments. The fixed point gets at just what is in common to the various ways a given piece of common knowledge can arise.

What about the iterate approach? We will show that for the relation of having information, the fixed-point approach is equivalent to the iterate approach, provided we restrict ourselves to finite situations. Without this assumption, though, the iterate approach, with only countably many iterations, is far too weak. In general, we must iterate on indefinitely into the transfinite.

Not only is the iterate approach too weak. When we move from having information to knowing, then even two iterations are unjustified. In general, the iterate approach is incomparable and really seems to miss the mark. We will see just what assumptions *are* needed to guarantee that the iterate account is equivalent to the fixed-point account.

Modeling Shared Information

In developing our model, we will follow the general line used in *The Liar*[11] in three ways. First, we take our metatheory to be ZF/AFA, a theory of sets that admits of circularity. We do this because ZF/AFA offers the most elegant mathematical setting we know for modeling circularity. Space does not permit us to give an introduction to this elegant set theory. We refer the reader to Chap. 3 of this book, or to Aczel's lectures[12] for an introduction.

Second, we follow the approach taken in *The Liar* in paying special attention to "situations," or "partial possible worlds." As far as this paper goes, the reader can think of a situation as simply representing an arbitrary set of basic facts, where a fact is simply some objects standing in some relation. Actually, in this paper, situations play a dual role. On the one hand they represent parts of the world. On the other hand they represent information about parts of the world. Thus, for example, we will define what it means for one situation s_0 to support another situation s_1, in the sense that s_0 contains enough facts to support all the facts in s_1.

[11] J. Barwise and J. Etchemendy, *The Liar: An Essay on Truth and Circularity* (New York: Oxford University Press, 1987).

[12] P. Aczel, *Non-well-founded Sets* (CSLI Lecture Notes (Chicago: University of Chicago Press, 1987 (to appear)).

Finally, on the trivial side, we also follow *The Liar* in considering a domain of card players as our domain to be modeled. We use this domain because it is simple, and because the existence of common knowledge is absolutely transparent to anyone who has ever played stud poker. And while the example is simple, there is enough complexity to illustrate many of the general points that need making. However, there is nothing about the results that depend on this assumption. You could replace the relation of having a given card with any relation whatsoever, and the results would still obtain.

Example 37.1 Simply by way of illustration, we have a running example, a game of stud poker. To make it very simple, we will use two card stud poker,[13] with two players, Claire and Max. We will assume that the players have the following cards:

Player	Down card	Up card
Claire	$A\spadesuit$	$3\clubsuit$
Max	$3\spadesuit$	$3\diamond$

Except for the rules and the idiosyncrasies of the other players, all the information available to the players is represented in this table. Note that based on what he sees, Max knows that he has the winning hand, or at least a tie, but Claire thinks she has a good chance of having the winning hand. The question before us is how best to model the informational difference between up cards and down cards.

Notice how different this situation would be from draw poker, where all cards are down, even if each player had cheated and learned the value of the second card. Anyone who has played poker will realize the vast difference. The reason is that in the standard case, the values of all the up cards is common knowledge, but in the second it isn't. Our aim, then, is to use tools from logic to model the three approaches to the common knowledge and shared information present in such a situation.

We reiterate that we use this simple card domain simply by way of making things concrete. We could equally well treat the more general case, if space permitted. We use S for the relation of seeing (or more generally of having information), H for the relation of having a card, and appropriate tuples to represent facts involving these relations. Thus, the fact that Max has the $3\spadesuit$ will be represented by the triple $\langle H, \text{Max}, 3\spadesuit\rangle$. The fact that Claire sees this will be represented by $\langle S, \text{Claire}, \{\langle H,$

[13]For the reader unfamiliar with two card stud poker, here is all you need to know to follow the example. First each player is dealt one card which only he is allowed to see, and there is a round of betting. Then each player is dealt one card face up on the table and there is another round of betting. Hands are ranked and players bet if they think their hand is best. But they can also drop out of the round at any point. After both rounds of betting are over, the hands are displayed, so that all players can see who won. As far as the ranking, all that matters is that a hand with a matching pair is better than a hand with no pairs. But among hands with no pairs, a hand with an ace is better than a hand with no ace.

Max, 3♠⟩}⟩. The question is how to adequately represent the common knowledge, or public information, of the up cards, like the fact ⟨*H*, Max, 3♢⟩ that Max has the 3♢. Thus for our formal development we have primitives: players $p_1, \ldots, p_n$, cards A♠, K♠, ..., 2♣, and relations *H* for the relation of having some card and *S* for the relation of seeing or otherwise having the information contained in some situation.

Comparing the Iterate and Fixed Point Accounts

Definition 37.1

1. The (models of) *situations* and *facts*[14] form the largest classes *SIT*, *FACT* such that:
   - $\sigma \in FACT$ iff σ is a triple, either of the form $\langle H, p, c\rangle$, where p is a player and c is a card, or of the form $\langle S, p, s\rangle$, where p is a player and $s \in SIT$.
   - A set s is in *SIT* iff $s \subseteq FACT$.
2. The *wellfounded situations* and *wellfounded facts* form the smallest classes *Wf-SIT* and *Wf-FACT* satisfying the above conditions.

Routine monotonicity considerations suffice to show that there are indeed largest and smallest such collections. If our working metatheory were ordinary ZF set theory, then these two definitions would collapse into a single one. However, working in ZF/AFA, there are many nonwellfounded situations and facts. A fact $\sigma = \langle R, a, b\rangle$ being in some situation s represents the fact of the relation R holding of the pair a, b in s, and is said to be a *fact of s*.

Example 37.1, Cont'd The basic situation s_0 about which player has which cards is represented by the following situation: $s_0 =$

$$\{\langle H, \text{Claire}, A\spadesuit\rangle\ ,\ \langle H, \text{Max}, 3\diamondsuit\rangle\ ,\ \langle H, \text{Claire}, 3\clubsuit\rangle\ ,\ \langle H, \text{Max}, 3\spadesuit\rangle\}$$

Abbreviations We sometimes write $(p_i\ Hc)$ for the fact $\langle H, p_i, c\rangle$, and similarly $(p_i\ Ss)$ for the fact $\langle S, p_i, s\rangle$. We write $(p_i\ S\sigma)$ for $(p_i\ Ss)$ where $s = \{\sigma\}$. All of our facts are atomic facts. However, our situations are like conjunctive facts. Hence we sometimes write $\sigma \wedge \tau$ for the situation $s = \{\sigma, \tau\}$, and so we can write $(p_i\ S\ (\sigma \wedge \tau))$ for $p_i\ Ss$, where $s = \{\sigma, \tau\}$. Similarly when there are more conjuncts.

Example 37.1, Cont'd With these tools and abbreviations, we can discuss the first two approaches to the public information about the up cards in our example. Toward this end, let $s_u =$

$$\{\langle H, \text{Claire}, 3\clubsuit\rangle\ ,\ \langle H, \text{Max}, 3\diamondsuit\rangle\}$$

which represents situation concerning the up cards.

[14] In order to keep this paper within bounds, I am restricting attention only to positive, nondisjunctive facts.

Iterates On this account, the fact that s_u is public information would be represented by an infinite number of distinct wellfounded facts: (Claire Ss_u), (Max Ss_u), Claire S(Claire Ss_u), (Max S(Claire Ss_u)), etc., in other words, by a wellfounded though infinite situation.

Fixed-Point On this account, the fact that s_u is publicly perceived by our players can be represented by the following public situation s_p:

$$s_p = \left\{\text{Claire } S\left(s_u \cup s_p\right), (\text{Max } S\left(s_u \cup s_p\right)\right\}$$

By contrast with the iterate approach, this situation contains just two facts. However, it is circular and so not wellfounded. The Solution Lemma of ZF/AFA guarantees that the sets used to represent the situation s_p exists.

It will be useful for later purposes to have a notation for some of the situations that play a role in our example. First, let the situations s_1, s_2 represent the visual situations, as seen by each of Claire and Max, respectively, including both the up cards and what each sees about what the others see. Consider also the larger situation s_w that represents the whole. Let $s_w = s_0$ (from above) union the set of the following facts:

$$\langle S, \text{Claire}, (\text{Claire } HA\spadesuit)\rangle, \langle S, \text{Max}, (\text{Max } H3\spadesuit)\rangle,$$
$$\langle S, \text{Claire}, s_1\rangle, \langle S, \text{Max}, s_2\rangle$$

where the first two facts represent what each player sees about his own down cards, and, e.g., s_1 is everything relevant seen by Claire, with facts s_u (= the "up" cards, as above) plus the fact (S, Max, s_2). Notice that s_1 is a constituent of s_2, and vice versa, so that s_w is a circular, nonwellfounded situation.

The next task is to define what it means for a fact a to hold in a situation s, which we write $s \models \sigma$, so that we can show that the situation s_w does satisfy the fixed point fact situation s_p defined above, as well as the above iterates.

Definition 37.2 The relation $\models$ is the largest subclass of $SIT \times FACT$ satisfying the following conditions:

- s $\models$ (pHc) iff $\langle H, p, c\rangle \in s$
- $s \models (pSs_0)$ iff there is an s_1 such that $\langle S, p, s_1\rangle \in s$, and for each $\sigma \in s_0$, $s_1 \models \sigma$.

The motivation for the second clause should be fairly obvious. If, in s, a player p sees (or otherwise has the information) s_1, and if s_1 satisfies each $\sigma \in s_0$, then in s that same player p sees (or otherwise has the information) s_0. This would not be a reasonable assumption about the usual notion of knowledge, since knowledge is not closed under logical entailment.

There is a difference with the possible worlds approach that sometimes seems puzzling to someone familiar with the traditional modal approach to knowledge. In p.w. semantics, partial situations are represented by the set of all possible worlds

compatible with them. As a result, whereas we can use an *existential* quantifier in clause (2) of our definition over situations about which p has information, the p.w. approach is forced to use a universal quantifier over possible worlds.

The reader can verify that all of the facts of s_p and the hierarchy of iterates of our running example indeed hold in the situation s_w. We also note that it follows from the definition that for all facts σ, if $\sigma \in s$ then $s \models \sigma$. However, the converse does not hold.

We extend our notation a bit and write $s_1 \models s_2$ provided $s_1 \models \sigma$ for each $\sigma \in s_2$.

As a companion of this notion of holding in, there is a notion of hereditary subsituation.[15] Intuitively, s_1 is a hereditary subsituation of s_2, written $s_1 \sqsubseteq s_2$, if all the information present in s_1 is present in s_2.

Definition 37.3 The hereditary subsituation relation $\sqsubseteq$ is the largest relation on $SIT \times SIT$ satisfying: $s_1 \sqsubseteq s_2$ iff:

- If $\langle H, p, c\rangle \in s_1$, then $\langle H, p, c\rangle \in s_2$;
- If $\langle S, p, s_0\rangle \in s_1$, then there is an s such that $s_0 \sqsubseteq s$ and $\langle S, p, s\rangle \in s_2$.

Proposition 37.1

1. *If* $s_1 \models \sigma$ *and* $s_1 \sqsubseteq s_2$, *then* $s_2 \models \sigma$.
2. *For all situations* s_0 *and* s_1, *the following are equivalent:*

   (a) $s_1 \sqsubseteq s_2$
   (b) $s_2 \models \sigma$ *for each* $\sigma \in s_1$.

Proof Limitations of space in this volume prevent us from doing more than hint at the proofs of the results in this paper. In this case we note that (1) is a simple consequence of the maximality of the $\models$ relation. Likewise, the implication from (2a) to (2b) is a consequence of the maximality of the $\models$ relation. The converse is a simple consequence of the maximality of the $\sqsubseteq$ relation. □

We say that situations s_0, s_1 are *informationally equivalent*, $s_0 \equiv s_1$, if the same facts hold in them. This is clearly an equivalence relation on situations. By the above lemma, $s_0 \equiv s_1$ if and only if each is a hereditary subsituation of the other. Distinct situations are often informationally equivalent. For example, suppose s_0' is a proper subset of the set s_1' of facts. Consider the situation $s_1 = \{\langle S, \text{Max}, s_1\rangle\}$, where Max has the information s_1, with the situation s_0 where there are two facts, that Max has the information s_0' and that he has the information s_1'. Using the fact just mentioned it is clear that $s_0 \equiv s_1$.

To compare the iterate and the fixed point approaches, we will show how an arbitrary fact θ (or situation s) gives rise to a transfinite sequence of wellfounded facts θ^α (or wellfounded situations s^α), for arbitrary ordinal α, finite or infinite. We use Tr for the conjunction of the empty situation, a fact that holds in every situation.

[15] In more recent joint work with Aczel, a generalization of this relation takes center stage.

Definition 37.4 The transfinite sequence $\langle \theta^\alpha \mid \alpha \in \mathit{Ordinals} \rangle$ of wellfounded facts associated with an arbitrary fact θ is defined by induction on ordinals as follows: for any θ, $\theta^0 = Tr$, and for $\alpha > 0$ we have:

$$(pHc)^\alpha = (pHc)$$
$$(pSs)^\alpha = (pS\ \ s^{<\alpha})$$

where

$$s^{<\alpha} = \left\{ \sigma^\beta \,\middle|\, \sigma \in s, \beta < \alpha \right\}$$

Similarly, for any situation s we define the transfinite sequence $\langle s^\alpha \mid \alpha \in \mathit{Ordinals} \rangle$ by letting $s^\alpha = \{\sigma^{\ \alpha} | \sigma \in s\}$.

The reader should verify that if we apply this definition to the fixed point fact in our example, we generate the iterates for all the finite ordinals, but then we go on beyond them into the transfinite.

We say that a fact σ *entails* a fact τ, written $\sigma \Rightarrow \tau$, if for every situation s, if $s \models \sigma$ then $s \models \tau$.

Theorem 37.2 Let θ be some fact.

1. *For all* α, $\theta \Rightarrow \theta^\alpha$.
2. *If each approximation fad* θ^α *holds in a situation* s, *then so does* θ.
3. *Assume that* κ *is a regular cardinal, and that* s *is a situation of size less than* κ. *If each approximation* θ^α, *for* $\alpha < k$, *holds in* s, *then so does* θ.

Proof The first is proved by means of a routine induction on α. The second is a consequence of the maximality of $\models$ and is not too difficult to prove. The third is a strengthening of the second involving routine cardinality considerations. □

Corollary 37.3 *Let* θ *be any fact, and let* s_w *be the set of all finite approximations of* θ. *Then, for any finite situation* s, $s \models \theta$ *iff* $s \models s_w$.

Refinement (3) of (2) of Theorem 37.2, and so the above corollary, were not present in the original working paper referred to above. They were discovered later in joint work with Peter Aczel. This result shows that the finite approximations of a circular fact will be equivalent to it, with respect to *finite* situations. This is a bit unsatisfactory, since the iterates themselves form an infinite situation. Still, it is the best we can hope for. However, in general, when we drop this restriction to finite models, one must look at the whole transfinite sequence of approximations. No initial segment is enough, as simple examples show. In this sense, the usual iterate approach is actually weaker than the simpler fixed-point approach.

When we move from having shared information to knowing, additional considerations must be brought to bear, as we will see below.

Comparing the Fixed Point and Shared Environment Approaches

To compare the shared environment approach with the fixed point approach, we introduce a simple second-order language which allows us to make existential claims about situations of just the kind made in the shared environment approach. We call the statements of this language ∃-statements. Before giving the definition, let's give an example. The following ∃-statement

$$\exists e\,[e = ((\text{Claire } H3\clubsuit) \wedge (\text{Claire } S\ \ e) \wedge (\text{Max } S\ \ e))]$$

is one shared environment analysis of the fact that Claire and Max share the information that Claire has the 3♣. Notice that what we have here is a simple, finite, wellfounded statement, but one that could only hold of nonwellfounded situations. Similarly, there is a fairly simple ∃-statement explicitly describing the situation s_w in our running example.

To define our language, we introduce variables $e_1, e_2, \ldots$ ranging over situations, in addition to constants for the cards and players. In fact, we do not bother to distinguish between a card or player and the constant used to denote it in statements. For atomic statements we have those of the form $(p_i Hc)$ (where P_i is a player and c is a card) and $(p_i Se_j)$. The set of ∃-statements forms the smallest set containing these atomic statements and closed under conjunction (∧), existential quantification over situations $(\exists e_j)$ and the rule: if Φ is a statement so is $(e_j \models \Phi)$. We are thus using ⊨ both for a relation symbol of our little language, as well as a symbol in our metalanguage. No more confusion should result from this than from the similar use of constants for cards and people. Finally, given any function f which assigns situations to variables, we define what it means for a statement Φ to hold in a situation s relative to f, written $s \models \Phi[f]$, in the expected way.

Definition 37.5

1. If Φ is an atomic statement, then $s \models \Phi[f]$ iff the appropriate fact is an element of s. In particular, if Φ is $(p_i Se_j)$, then $s \models \Phi[f]$ iff $\langle S, p_i, f(e_j)\rangle \in s$.
2. If Φ is $\Phi_1 \wedge \Phi_2$ then $s \models \Phi[f]$ iff $s \models \Phi_1[f]$ and $s \models \Phi_2[f]$
3. If Φ is $\exists e_j \Phi_0$ then $s \models \Phi[f]$ iff there is a situation s_j so that $s \models \Phi_0[f(e_j/s_j)]$
4. If Φ is $(e_j \models \Phi_0)$ then $s \models \Phi[f]$ iff the situation $s_j = f(e_j)$ satisfies $s_j \models \Phi_0[f]$.

A *closed* ∃-statement is one with no free variables, as usual. If Φ is closed, we write $s \models \Phi$ if some (equivalently, every) assignment f satisfies $s \models \Phi[f]$.

Notice that the ∃-statements are all finite and wellfounded. (The results that follow would hold equally well if we allowed infinite conjunctions and infinite strings of quantifiers, except for the word "finite" in Theorem 37.5 below.) Nevertheless, some of them can only hold of nonwellfounded situations, as the above example shows.

We want to show that any ∃-statement can be approximated in a certain sense by a fixed point situation. In particular, if we take as our ∃-statement one that expresses a shared environment approach to shared information, the resulting situation will be the one that characterizes the fixed point approach. Then, using the transfinite wellfounded iterates approximating the fixed point approach, we obtain a transfinite sequence of wellfounded facts approximating any ∃-statement.

Let us say that a situation s_Φ almost characterizes the ∃-statement Φ if $s_\Phi \models \Phi$ and for every situation $s \models \Phi$, we have $s \models s_\Phi$. For example, if we take our above example of an ∃-statement, then the following situation can easily be seen to almost characterize it:

$$s = \{\langle H, \mathit{Claire}, 3\clubsuit\rangle, \langle S, \mathit{Claire}, s\rangle, \langle S, \mathrm{Max}, s\rangle\}$$

Clearly our statement is true in this model. It is also easy to see that s is a hereditary subsituation of any situation which is a model of our statement, so by Proposition 37.1, s almost characterizes the statement. This definition is justified by the following result, which is an easy consequence of Proposition 37.1.

Proposition 37.4 *Suppose that the situation s almost characterizes the* ∃*-statement* Φ. *Then for any fact* σ, *the following are equivalent:*

1. σ *is entailed by* Φ, *i.e.*, σ *holds in all models of* Φ
2. $s \models \sigma$

The following is the main result of this paper. It shows the extent to which the shared environment approach can be approximated by the fixed point and iterate approaches.

Theorem 37.5 *Every* ∃*-statement* Φ *is almost characterized by some finite situation* s_Φ.

Proof First one establishes a normal form lemma for ∃-statements, where all the quantifiers are pulled out front. One then uses the Solution Lemma of AFA to define the desired situation. The proof that it almost characterizes the statement uses Proposition 37.1. □

However, there is a distinct sense in which ∃-statements are more discriminating than the situations that almost characterize them. For example, compare our above example of an ∃-statement with the following:

$$\exists e_1, e_2\,[e_1 \models ((\text{Claire } H3\clubsuit) \wedge (\text{Claire } S\ \ e_2)) \wedge e_2 \models ((\text{Claire } H3\clubsuit)\, J \wedge (\text{Max } S\ \ e_1))]$$

Clearly any model of our first statement is a model of our second. However, it is easy to see that there are models of our second that are not models of our first. (Think of a case where the card is not an up card, but is down, but where there are suitably placed mirrors.) On the other hand, these two statements are almost characterized by exactly the same situations. Or, in view of Proposition 37.4, the two statements entail the same facts, both wellfounded and circular.

Intuitively, what is going on here is that both of these statements represent ways in which Max and Claire might share the information that Claire has the 3♣. The first would be the one predicted by a literal reading of the Clark and Marshall account, but the second is clear in the spirit of that account. However, this means that since they are not equivalent, neither one can be the right characterization of the shared information. Rather, what they represent are two distinct ways, among many, that Max and Claire might have come to have the shared information. We leave it to the reader to work out analogous inequivalent ∃-statements that also give rise to the shared information in our running example.

We conclude this section by observing that the results can be extended to the case where we allow disjunctions to occur in ∃-statements, if one also allows disjunctive facts.

Conclusions

In thinking about shared information and common knowledge, it is important to keep three questions separate: (i) What is the correct analysis of common knowledge? (ii) Where does it come from? (iii) How is it used?

It would be neat if these three questions got their answers from the three different approaches in the literature. The results discussed above prompt us to propose that the fixed-point approach is the right analysis of the notion, and that it typically arises through some sort of shared environment.

However, by definition, the epistemically neutral case we have been studying is divorced from questions of use. To think about how shared information gets used, we turn to the epistemic case. Let us suppose that the fixed-point approach, or something like it, characterizes common knowledge, and the shared-environment approach characterizes the way in which common knowledge commonly arises. Does it follow that the iterate approach approximates common knowledge, or perhaps how it is used?

It seems that it can't. A clear difference between having information and knowing arises in the respective relationships between the fixed-point facts and its approximations. In the nonepistemic case, it is a matter of logical entailment. However, in the latter case, the fixed-point fact will simply not entail the analogous approximations. To see why, let's consider an example.

Example 37.2 Consider the following situation s, where we use K for the relation of knowing of a situation:

$$\langle H, \text{Max}, 3\diamondsuit\rangle\ ,\ \langle K, \text{Claire}, s\rangle\ ,\ \langle K, \text{Dana}, s\rangle\ ,\ \langle K, \text{Max}, s\rangle$$

It seems clear that the fact

$$\theta = (\text{Max}\ H3\diamondsuit) \wedge (\text{Claire}\ K\theta) \wedge (\text{Dana}\ K\theta) \wedge (\text{Max}\ K\theta)$$

holds in this situation. However, is it a fact in this situation that, say, Max knows that Dana knows that Claire knows that he, Max, has the $3\diamond$? And even more iterations?

It seems clear that it will not in general be true. After all, some sort of inference is required to get each iteration, and the players might not make the inference. They are, after all, only 3 years old. And even if Claire makes her inference, Dana may have legitimate doubts about whether Claire has made her inference. But once one player has the least doubt about some other player's making the relevant inference, the iterated knowledge facts breaks down. That is, once the making of an inference is implausible, or even just in doubt, the next fact in the hierarchy is not really a *fact* at all.

It is usually said that the iterate account assumes that all the agents are perfectly rational, that is, that they are perfect reasoners. This example also shows that it in fact assumes more: it assumes that it is *common knowledge* among the agents that they are all perfectly rational. It is only by making this radical idealization, plus restricting attention to finite situations, that the iterate account is equivalent to the fixed-point account. And the idealization requires the very notion that one is trying to understand in the first place.

We began this section by asking three questions. We have proposed answers to the last two of them, and suggested that the third question, about how common knowledge is used, is not answered by the iterate approach. But then how *do* people make use of common knowledge in ordinary situations?

My own guess is that common knowledge per se, the notion captured by the fixed-point analysis, is not actually all that useful. It is a necessary but not a sufficient condition for action. What suffices in order for common knowledge to be useful is that it arise in some fairly straightforward shared situation. The reason this is useful is that such shared situations provide a basis for perceivable situated action, action that then produces further shared situations. That is, what makes a shared environment work is not just that it gives rise to common knowledge, but also that it provides a stage for maintaining common knowledge through the maintainence of a shared environment. This seems to me to be part of the moral of the exciting work of Parikh, applying ideas of game theory to the study of communication.[16]

It seems to me that the consequences of this view of common knowledge are startling, if applied to real world examples, things like deterrence (mutual assured destruction, say). Indeed, it suggests a strategy of openness that is the antithesis of the one actually employed. But that goes well beyond the scope of this conference.

Finally, let me note that the results here do not lend themselves to an immediate comparison with other mathematical models of common knowledge, especially the approaches in game theory. It would be interesting to see a similar analysis there, one that pinpoints the finiteness or compactness assumption that must be lurking behind the Tan and Ribeiro da Cost Werlang result.

[16]Prashant Parikh, "Language and strategic inference," Ph.D. Dissertation, Stanford University, 1987.

Chapter 38
The Logic of Public Announcements, Common Knowledge, and Private Suspicions

Alexandru Baltag, Lawrence S. Moss, and Sławomir Solecki

Introduction: Example Scenarios and Their Representations

We introduce the issues in this paper by presenting a few *epistemic scenarios*. These are all based on the Muddy Children scenario, well-known from the literature on knowledge. The intention is to expose the problems that we wish to address. These problems are first of all to get models which are faithful to our intuitions, and then to build and study logical systems which capture some of what is going on in the scenarios.

The cast of characters consists of three children: A, B, and C. So that we can use pronouns for them in the sequel, we assume that A is male, and B and C are female. Furthermore, A and B are dirty, and C is clean. Each of the children can see all and only the others. It is known to all (say, as a result of a shout from one of the parents) that at least one child is dirty. Furthermore, each child must try to figure out his or her state only by stating "I know whether I'm dirty or not" or "I don't know whether I'm dirty or not." They must tell the truth, and they are perfect reasoners in the sense that they know all of the semantic consequences of their knowledge. The opening situation and these rules are all assumed to be common knowledge.

Scenario 1. After reflection, A and B announce to everyone that at that point they do not know whether they are dirty or not. (The reason we are having A and B make this announcement rather than all three children is that it fits in better with our scenarios to follow.) Let α denote this announcement.

A. Baltag (✉)
ILLC, University of Amsterdam, The Netherlands
e-mail: thealexandrubaltag@gmail.com

L.S. Moss • S. Solecki
Mathematics Department, Indiana University, Bloomington, IN 47401, USA
e-mail: lsm@cs.indiana.edu; ssolecki@indiana.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_38

As in the classical Muddy Children, there are intuitions about knowledge before and after α. Here are some of those intuitions. Before α, nobody should know that he or she is dirty. However, A should think that it is possible that B knows. (For if A were clean, B would infer that she must the dirty one.) After α, A and B should each know that they are dirty, and hence they know whether they are dirty or not. On the other hand, C should not know whether she is dirty or not.

Scenario 1.5. This scenario begins after α. At this point, A and B announce to all three that they *do* know whether or not they are dirty. We'll call this event α'. Our intuition is that after α', C should know that she is not dirty. Moreover, A and B should know that C knows this. Actually, the dirty-or-not states of all the children should be *common knowledge* to all three.

Scenario 2. As an alternative to the first scenario, let's assume that C falls asleep for a minute. During this time, A and B got together and told each other that they didn't know whether they were dirty or not. Let β denote this event. After β, C wakes up. Part of what we mean by β is that C does not even consider it possible that β occurred, and that it's common knowledge to A and B that this is the case. Then our intuitions are that after β, C should "know" (actually: believe) that A does not know whether he is dirty (and similarly for B); and this fact about C is common knowledge for all three children. Of course, it should also be common knowledge to A and B that they are dirty.

Scenario 2.5. Following Scenario 2, we again have α': A and B announce that they do know whether they are dirty or not. Our intuitions are not entirely clear at this point. Surely C should suspect some kind of cheating or miscalculation on the part of the others. However, we will not have much to say about the workings of this kind of real-world sensibility. Our goal will be more in the direction of modeling different alternatives.

Scenario 3. Now we vary Scenario 2. C merely feigned sleep and thought she heard both A and B whispering. C cannot be sure of this, however, and also entertains the possibility that nothing was communicated. (In reality, A and B did communicate.) A and B for their part, still believe that C was sleeping. We call this event γ.

One might at first glance think that A and B's "knowledge" of C's epistemic state is unchanged by γ. After all, the communication was not about C. However, we work with a semantic notion of knowledge, and after γ, A and B know that they are dirty, hence then know that C knows that they are dirty. A and B did not know this at the outset.

So we need to revise the initial intuition. What is correct is that if C knows some fact φ before γ, then after γ, A and B know (or rather, believe) that C knows φ. This is because after γ, A and B not only know the clean-or-dirty state of everyone, they (therefore) also know exactly which possibilities everyone is aware of, which they discard as impossible, etc. So each of them can reconstruct C's entire epistemic state. They believe that their reconstruction is current, but of course, what they reconstruct is C's original one, before γ.

Conversely, if after γ, A and B "know" that C knows φ, then before γ, C really did know φ. That is, the reconstruction is accurate. For example, after γ, A believes that C should not consider it possible that A knows that he is dirty. However, C thinks it is possible that A knows he is dirty.

There is a stronger statement that is true: C knows φ before γ iff after γ, it is common knowledge to A and B that each of them knows that C knows φ. Intuitively, this hold because each of A and B knows that both of them are able to carry out the reconstruction of C's state.

Our final intuition is that after γ, C should know that if A were to subsequently announce that he knows that he is dirty, then C would know that B knows that she is dirty.

Scenario 3.5. Again, continue Scenario 3 by α'. At this point, C should know that her suspicions were confirmed, and hence that she is not dirty. For their part, A and B should think that C is confused by α': they should think that C is as she was following Scenario 2.5.

Scenario 4. A and B are on one side of the table and C is on the other, dozing. C wakes up at what looks to her like the middle of a joint confession by A and B. The two sides stare each other down. In fact, A and B have already communicated. We call this action δ. So C suspects that δ is what happened, but can't tell if it was δ or nothing. For their part, A and B see that C suspects but does not know that δ happened.

The basic intuition is that after δ, it should be common knowledge to all three that C suspects that the communication happened. Even if C thinks that A and B did not communicate, C should not think that she is sure of this.

One related intuition is that after δ, it should be common knowledge that C suspects that *A knows that he is dirty.* As it happens, this intuition is wrong. Here is a detailed analysis: C thinks it possible that everyone is dirty at the outset, and if this were the case then the announcement of B's ignorance would not help A to learn that he is dirty; from A's point of view, he still could be clean and B would not know that she is dirty. C's view on this does not change as a result of δ, so afterwards, C still thinks that it could be the case that A says, "It's possible that B and C are the dirty one and I am clean, Hence C would see my clean face and not suspect that I know that I am dirty." So it certainly should not be common knowledge that C suspects that A knows he is dirty.

Notice also that C would say after δ: "I think it is possible that no announcement occurred, and yet A thinks it possible that B is the only dirty one. In that case, what A would think that I suspect that A told B that he knows that he is *not* dirty. Of course, this is not what I actually suspect." The point is that C's reasoning about A and B's reasoning about her involves suspicion of a different announcement than we at first considered.

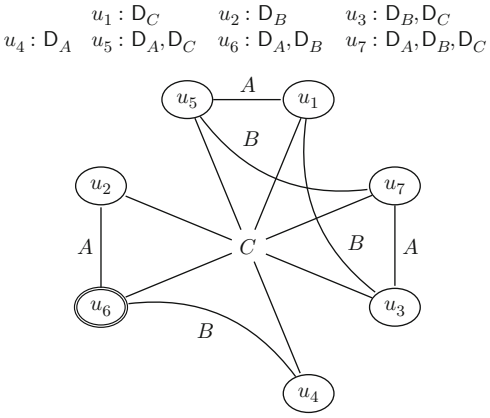
Scenario 4.5. Once again, we continue with α'. Our intuition is that this is tantamount to an admission of private communication by A and B. If we disregard

this and only look at higher order knowledge concerning who is and is not dirty, we expect that the epistemic state after α' is the same for all three children as it is at the end of Scenario 1.5.

Models

Now that we have detailed a few scenarios and our intuitions about them, it is time to construct some Kripke models as representations for them.

The Models *U and V.* We begin with a representation of the situation before α. We take the Kripke model U whose worlds are $u_1, \ldots, u_7$ and whose atomic information and structure is given below:



Note that we have not indicated any direction on the edges; they are all intended to be bidirectional. In addition, we have not shown self-loops for the agents. However, we intend in this model that all of the self-loops be present on all nodes for all agents.

As an example of reading the picture, in world u_3, A is clean, but B and C are dirty. Also, the worlds which A thinks are possible are u_3 and u_7. Thus, A sees that B and C are dirty, so A infers that the world is either u_3 or u_7. The lines for C are intended to go across the middle. The rest of the structure is explained similarly, except for the doubled ellipse around u_6. This specifies u_6 as the actual world in the model, the one which corresponds to our description of the model before α. Note that U incorporates some of the conventions stated in Scenario 1. For example, in each world, each child has a complete and correct assessment of which worlds are possible for all three reasoners.

Each of our intuitions about knowledge before α turns into a statement in the modal logic of knowledge. This logic has atomic sentences D_A, D_B, and D_C standing for "A is dirty", etc.; it has knowledge operators $\Box_A$, $\Box_B$, and $\Box_C$ along with the

usual boolean connectives. We are going to use the standard Kripke semantics for multi-modal logic throughout this paper. So given a model-world pair, say $\langle A, a\rangle$, and some agent, say D, we'll write

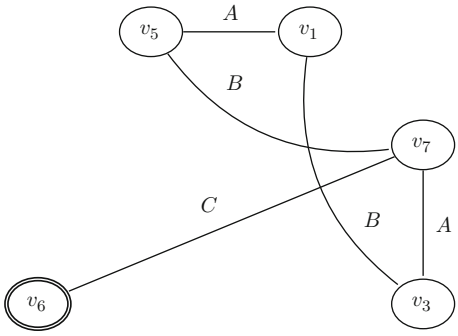
$$\langle K, k\rangle \models \Box_D \varphi \quad \text{iff} \quad \text{whenever } k \xrightarrow{D} l \text{ in } K, \text{ we have } \langle K, l\rangle \models \varphi.$$

The boolean connectives will be interpreted classically. We can then check the following:

$$\langle U, u_6\rangle \models \neg\Box_A \mathsf{D}_A \wedge \neg\Box_A\neg\mathsf{D}_A \wedge \neg\Box_B \mathsf{D}_B \wedge \neg\Box_B\neg\mathsf{D}_B \wedge \neg\Box_C \mathsf{D}_C \wedge \neg\Box_C\neg\mathsf{D}_C$$
$$\langle U, u_6\rangle \models \Diamond_A\Box_B \mathsf{D}_B$$

The model after α is the Kripke model V, shown below:

$$v_1 : \mathsf{D}_C \; v_3 : \mathsf{D}_B, \mathsf{D}_C \; v_5 : \mathsf{D}_A, \mathsf{D}_C \; v_6 : \mathsf{D}_A, \mathsf{D}_B \; v_7 : \mathsf{D}_A, \mathsf{D}_B, \mathsf{D}_C$$



The way we got V from U was to discard the worlds u_2 and u_4 of U, since in U at each of those worlds, either A or B would know if they were dirty. We also changed the u's to v's to avoid confusion, and to stress the fact that we get a new model. Turning back to our intuitions, we can see that the following holds:

$$\langle V, v_6\rangle \models \Box_A \mathsf{D}_A \wedge \Box_B \mathsf{D}_B \wedge \neg(\Box_C \mathsf{D}_C \vee \Box\neg\mathsf{D}_C)$$

The Model *W*. Scenario 1.5 elaborates Scenario 1 by the event α'. So we discard the worlds where this is false in V, and we obtain a one-world model $W = \{w_6\}$. In w_6, all three agents are dirty. The picture is

(We have renamed v_6 to w_6. We continue to omit the three self-loops, but shortly our pictures will begin to incorporate those when they are appropriate.) This model reflects our intuition that at this point, C should know that she is not dirty.

The Model *X*. This corresponds to Scenario 2. We start with U and see the effect of the private announcement β. The resulting model X is too large to show in a small diagram, so instead we use a chart:

World	A	B	C	$\xrightarrow{A}$	$\xrightarrow{B}$	$\xrightarrow{C}$
$u_1, \ldots, u_7$						
x_1			•	x_1, x_5	x_1, x_3	u_1
x_3		•	•	x_3, x_7	x_1, x_3	u_2, u_3
x_5	•		•	x_1, x_5	x_5, x_7	u_4, u_5
$x_6\surd$	•	•		x_6	x_6	u_6, u_7
x_7	•	•	•	x_3, x_7	x_5, x_7	u_6, u_7

The "real world" is x_6. Notice that the worlds $u_1, \ldots, u_7$ are also worlds in X. We did not put any information in the chart above for those worlds since it should be exactly the same as in U above. The reason for having these "old worlds" in X is that since C was asleep, the worlds that C considers possible after β should be just the ones that were possible before β. We can check that

$$\langle X, x_6\rangle \models \neg\Box_C(\Box_A \mathsf{D}_A \vee \Box_A \neg \mathsf{D}_A).$$

Let φ be the sentence above. Then also, $\langle X, x_6\rangle \models \Box^*_{\{A,B,C\}}\varphi$. This is our formal statement that it is common knowledge in the group of three children that φ holds. The semantics of this is that for all sequences $D_1, \ldots, D_m \in \{A, B, C\}^*$, $\langle X, x_6\rangle \models \Box_{D_1} \cdots \Box_{D_m}\varphi$. Note that we have no way of saying in the modal language that C suspects that an announcement happened; the best we can do is (roughly) to say that C thinks that some sentence ψ is possible in the sense that ψ holds in some possible world. Of course, we have no way to say that A and B know that C was asleep, either.

Note as well that in X, we do not have $x_6 \xrightarrow{C} x_6$. In other words, the real world would not be possible for C. This is some indication that something strange is going on in this model. Further, we consider the model of what happens after A and B's announcement. Then in this model, *no worlds* would be accessible for C from the actual world. These anomalies should justify our interest in the more complicated scenarios and models involving suspicions of announcements.

The Model obtained by announcing α' in *X*. This would be the model with one world, say x_6^* where A and B are dirty, and whose structure is given by

We have not only deleted the worlds where either A or B does not know that they are dirty in X, but we also discarded all worlds not reachable from the new version x_6^* of x_6. The anomaly here is that C thinks no worlds are possible.

World	A	B	C	$\xrightarrow{A}$	$\xrightarrow{B}$	$\xrightarrow{C}$
$u_1, \ldots, u_7$						
x_1, x_3, x_5, x_6, x_7						
y_1			•	x_1, x_5	x_1, x_3	y_1, y_1'
y_3		•	•	x_3, x_7	x_1, x_3	y_3, y_2', y_3'
y_5	•		•	x_1, x_5	x_5, x_7	y_5, y_4', y_5'
$y_6\surd$	•	•		x_6	x_6	y_6, y_7, y_6', y_7'
y_7	•	•	•	x_3, x_7	x_5, x_7	y_6, y_7, y_6', y_7'
y_1'			•	u_1, u_5	u_1, u_3	y_1, y_1'
y_2'		•		u_2, u_6	u_2	y_3, y_2', y_3'
y_3'		•	•	u_3, u_7	y_1, y_3	y_3, y_2', y_3'
y_4'	•			u_4	u_4, u_6	y_5, y_4, y_5'
y_5'	•		•	u_1, u_5	u_5, u_7	y_5, y_4', y_5'
y_6'	•	•		u_2, u_6	u_4, u_6	y_6, y_7, y_6', y_7'
y_7'	•	•	•	u_3, u_7	u_5, u_7	y_6, y_7, y_6', y_7'

Fig. 38.1 The model Y

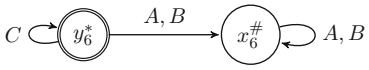
The Model *Y*. We consider γ from Scenario 3, in which C thought she might have heard A and B, while A and B think that C is unaware of γ. We get the model Y displayed in Fig. 38.1 above. Y has 24 worlds, and so we won't justify all of them individually. We will give a more principled construction of Y from W and γ, once we have settle on a mathematical model of γ. For now, the ideas are that the y worlds are those where the announcement happened, and the y' worlds are those in which it did not. Note that some of the y worlds are missing, since the truthful announcement by A and B presupposes that they don't know whether they are dirty in U at the corresponding world. The x's and u's are from above, and they inherit the accessibility relations which we have seen.

Now our main intuition here is that $\langle U, u_6\rangle \models \Box_C\varphi$ iff $\langle Y, y_6\rangle \models \Box^+_{\{A,B\}}\Box_C\varphi$. (The sentence $\Box^+_{\{A,B\}}\chi$ means that A knows φ, A knows B knows χ, etc. It differs from $\Box^*_{\{A,B\}}\chi$ in that it does not entail that χ is true.) To see this, note that $u_6 \xrightarrow{C} u_6, u_7$ and no other worlds. And the only worlds reachable from y_6 using one or more $\xrightarrow{A}$ or $\xrightarrow{B}$ transitions followed by a $\xrightarrow{C}$ transition are again u_6 and u_7.

Another intuition is that in $\langle Y, y_6\rangle$, C should think that it is possible that A knows that he is dirty. This is justified since $y_6 \xrightarrow{C} x_6$, and $\langle Y, x_6\rangle \cong \langle X, x_6\rangle$ (that is, the submodels of X and Y generated by x_6 are isomorphic), and $\langle X, x_6\rangle \models \Box_A\mathsf{D}_A$.

Our final intuition is that in $\langle Y, y_6\rangle$, C should know that if A were to subsequently announce that he knows that he is dirty, then C would know that B knows that she is dirty. To check this, we need to modify Y by deleting the worlds where A does not know that he is dirty. These include y_7, y_6' and y_7'. In the updated model, the only world accessible for C from (the new version of) y_6 is y_6 itself, and at y_6 in the new structure, B correctly knows she is dirty.

The Model obtained by announcing α' in Y. As when α' is announced in X, we only keep the worlds of Y worlds where both A or B do know they are dirty. So we drop y_7, y'_6, and y'_7.



We also only keep the worlds accessible from y_6 (this change is harmless). C knows she is not dirty. Technically, A and B "know" this, but this is for the nonsensical reason that they "know" that C knows everything.

The Model Z. Corresponding to Scenario 4, we get the model Z shown below.

w	A	B	C	$\xrightarrow{A}$	$\xrightarrow{B}$	$\xrightarrow{C}$
z_1			•	z_1, z_5	z_1, z_3	z_1, z'_1
z_2		•		z_2	z_2	z_2, z_3, z'_2, z'_3
z_3		•	•	z_3, z_7	z_1, z_3	z_2, z_3, z'_2, z'_3
z_4	•			z_4	z_4	z_4, z_5, z'_4, z'_5
z_5	•		•	z_1, z_5	z_5, z_7	z_4, z_5, z'_4, z'_5
z_6 √	•	•		z_6	z_6	z_6, z_7, z'_6, z'_7
z_7	•	•	•	z_3, z_7	z_5, z_7	z_6, z_7, z'_6, z'_7

w	A	B	C	$\xrightarrow{A}$	$\xrightarrow{B}$	$\xrightarrow{C}$
z'_1			•	z'_1, z'_5	z'_1, z'_3	z_1, z'_1
z'_2		•		z'_2, z'_6	z_2, z'_2	z_2, z_3, z'_2, z'_3
z'_3		•	•	z'_3, z'_7	z'_1, z'_3	z_2, z_3, z'_2, z'_3
z'_4	•			z'_4, z'_6	z'_4, z'_6	z_4, z_5, z'_4, z'_5
z'_5	•		•	z'_1, z'_5	z'_5, z'_7	z_4, z_5, z'_4, z'_5
z'_6	•	•		z'_2, z'_6	z'_4, z'_6	z_6, z_7, z'_6, z'_7
z'_7	•	•	•	z'_3, z'_7	z'_5, z'_7	z_6, z_7, z'_6, z'_7

Recall our last point in Scenario 4, that we need to consider a few possible announcements for C to suspect. This is reflected in the fact that the z worlds are of three types. In z_2, B announced that she knows whether she is dirty, and A announced that he doesn't. Similar remarks apply to z_4. In all other z worlds, both announced that they do not know. The worlds accessible from each of these is based on the relevant announcement. For example, in z_2, neither A nor B thinks any other world is possible. (One might think that $z_2 \xrightarrow{A} z_6$. But in z_6, B could not announce that she knows she is dirty. So if the world were z_2 and the relevant announcement made, then A would not think z_6 is possible.) The z' worlds are those in which no announcement actually happened.

Our key intuition was that it is common knowledge that C suspects that δ happened. This will not correspond to anything in the formal language $\mathcal{L}([\alpha], \Box^*)$ introduced later in this paper. (However, it will be representable in an auxiliary language about actions; see Example 3.) Informally, the intuition is valid for Z because for every z_i (or z'_i) there is some z_j (unprimed) such that $z_i \xrightarrow{C} z_j$ (or $z'_i \xrightarrow{C} z_j$). In addition, in this particular model there is a sentence in our formal language which happens to hold only at the worlds where an announcement occurred. Here is one:

$$\chi \quad \equiv \quad \Box_A \mathsf{D}_A \ \vee \ \Box_A \neg \mathsf{D}_A \vee \ \Diamond^*_{\{A,B\}} \Diamond_C \Box_A \mathsf{D}_A$$

So $\langle Z, z_6 \rangle \models \Box^*_{\{A,B,C\}} \Diamond_C \chi$.

The explanation of the mistaken intuition in Scenario 4 is that $z_6 \xrightarrow{C} z_7 \xrightarrow{A} z_3$, and $\Diamond_A \Box_A \mathsf{D}_A$ fails in z_2, z_3, z_2', and z_3'. Overall, $\langle Z, z_6\rangle \models \neg\Box_C\Box_A \Diamond_A \Box_A \mathsf{D}_A$.

The point that C's suspicion varies corresponds to the fact that $\Diamond_C\Diamond_A\Diamond_C\Box_A\neg\mathsf{D}_A$ holds at $\langle Z, z_6\rangle$. Indeed $z_6 \xrightarrow{C} z_6' \xrightarrow{A} z_2' \xrightarrow{C} z_2$, and $\langle Z, z_2\rangle \models \Box_A\neg\mathsf{D}_A$.

A few more involved statements are true in Z. For example, $\Box_{\{A,B,C\}^*}(\Box_A\mathsf{D}_A \rightarrow \Diamond_C\Box_A\mathsf{D}_A)$. It is common knowledge to all three that *if* A knows he is dirty, then C thinks it possible that A knows this.

The Model obtained by announcing α' in Z. This model is W from above. (Actually, it is bisimilar to W; see section "The larger language $\mathcal{L}([\alpha], \Box^*)$".) This corresponds to the intuition that Scenarios 2.5 and 4.5 lead to the same model.

Epistemic Actions

We will formalize a language in section "A Logical Language with Epistemic Actions" along with the notions of (epistemic) *action structure* and *actions*. Before we do that, it makes sense to present the idea informally based on the examples which we have already dealt with.

α and α': announcements to everyone. We first consider α of Scenario 1. Let ψ be given by

$$\psi \quad := \quad \neg(\Box_A\mathsf{D}_A \vee \Box_A\neg\mathsf{D}_A) \wedge \neg(\Box_B\mathsf{D}_B \vee \Box_B\neg\mathsf{D}_B) \tag{38.1}$$

So ψ says that neither A nor B know whether or not they are dirty. This is the *precondition* of the announcement, but it is not the *structure*. The structure of this announcement is quite simple (so much so that the reader will need to read further to get an idea for what we mean by structure). It is the following Kripke structure K: we take one point, call it k, and we set $k \xrightarrow{D} k$ for all $D \in \{A, B, C\}$. We call $\langle K, k\rangle$ an *action structure*. Along with K, we also have a *precondition*; this will be ψ from (38.1). To deal with action structures with more than one point, the precondition will be a function PRE from worlds to sentences. In this case, the function PRE is just $\{\langle k, \psi\rangle\}$. The tuple $\langle K, k, \text{PRE}\rangle$ will be an example of what we call an *action*. This particular action is our model of the announcement α. Henceforth we use the symbol α to refer ambiguously to the pretheoretic notion of the announcement event and to our mathematical model of it.

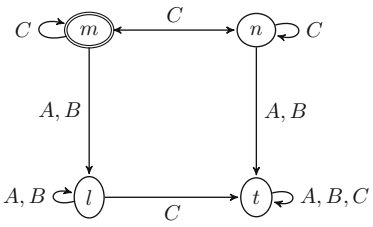
Another example of an announcement to everyone is α'. Here we just change ψ from (38.1) to the sentence ψ' which says that both A and B know whether or not they are dirty. Yet another example is the *null announcement*. This models the announcement of a tautology *true* to everyone. We'll write this as τ.

***β*: a secure announcement to a set of agents.** Next, suppose we have an announcement made to some possibly proper subset $\mathcal{B} \subseteq \mathcal{A}$ in the manner of Scenario 2. So there is some dispute as to what happened: the agents in $\mathcal{B}$ think that

there was an announcement, while those out of $\mathcal{B}$ are sure that nothing happened. We formalize this with a Kripke structure of two points, l and t. We set $l \xrightarrow{D} l$ for all $D \in \mathcal{B}$, $l \xrightarrow{D} t$ for $D \notin \mathcal{B}$, and $t \xrightarrow{D} t$ for all D. The point is that l here is the actual announcement, and the agents in $\mathcal{B}$ know that this is the announcement. The agents not in $\mathcal{B}$ think that t is for sure the only possible action, and t in this model will behave just like the null announcement. The precondition function will be called PRE in all of our examples. Here PRE is given by $\text{PRE}(l) = \psi$ and $\text{PRE}(t) = \mathit{true}$, where ψ is from (38.1). The action overall is $\langle L, l, \text{PRE}\rangle$, where $L = \{l, t\}$. We call this action β.

γ: an announcement with a suspicious outsider. This is based on Scenario 3. The associated structure has four points, as shown below:

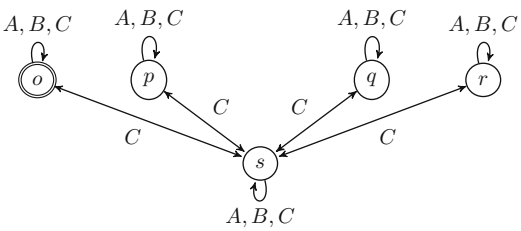
$$\text{PRE}(m) = \psi \;\; \text{PRE}(n) = \mathit{true} \;\; \text{PRE}(l) = \psi \;\; \text{PRE}(n) = \mathit{true}$$



The idea is that m is the (private) announcement that C suspects, and n is other announcement that C thinks is possible (where nothing was communicated by A and B). Then if m happened, A and B were sure that what happened was l; similarly, if n happened, A and B would think that t was what happened. We call this action γ; technically it is $\langle\{m, n, l, t\}, m, \text{PRE}\rangle$. We get a different action, say γ' if we use the same model as above but change the designated ("real") world from m to n.

δ: an announcement with common knowledge of suspicion. Corresponding to Scenario 4, we have the following model. In it ψ_A denotes the sentence saying that A knows whether he is dirty but B does not, ψ_B the sentence saying that B knows whether she is dirty but A does not, and $\psi_\emptyset$ the sentence stating that neither knows.

$$\text{PRE}(o) = \psi \;\; \text{PRE}(p) = \psi_A \;\; \text{PRE}(q) = \psi_B \;\; \text{PRE}(r) = \psi_\emptyset \;\; \text{PRE}(s) = \mathit{true}$$



We call this action δ. There are five possible actions here, depending on whether it was ψ, ψ_A, ψ_B, $\psi_\emptyset$ or nothing which was announced. In each case, A and B are sure of what happened. Even if nothing actually happened (s), C would suspect one of the other four possibilities. In those, C still considers it possible that nothing happened.

Still to come. The reader is perhaps wondering what the actual connection is between the (formal) actions just introduced and the concrete models of the previous section. The connection is that there is a way of taking a model and an action and producing another model. When applied to the specific model U and the actions of this section, we get the models $V, \ldots, Z$. We delay this connection until section "Semantics" below, since it is high time that we introduce our language of epistemic actions and its semantics. The point is that there is a principled reason behind the models.

The question also arises as to whether there are any principles behind the particular actions which we put down in this section. As it happens, there is more which can be said on this matter. We postpone that discussion until section "More on Actions", after we have formally defined the syntax and semantics of our logical languages.

The Issues

The main issue we address in this paper is to formally represent *epistemic updates*, i.e., changes in the information states of agents in a distributed system. We think of these changes as being induced by specific information-updating actions, which can be of various types: (1) information-gathering and processing (e.g., realizing the possibility of other agents' hidden actions, and more generally, learning of any kind of new possibility via experiment, computation, or introspection); (2) information-exchange and communication (learning by sending/receiving messages, public announcements, secret interception of messages, etc.); (3) information-hiding (lying or other forms of deceiving actions, such as communication over secret channels, sending encrypted messages, holding secret suspicions); (4) information-loss and misinformation (being lied to, starting to have gratuitous suspicions, non-introspective learning, wrong computations or faulty observations, paranoia); (5) and more generally sequential or synchronous combinations of all of the above.

Special cases of our logic, dealing only with public or semi-public *announcements to mutually isolated groups*, have been considered in Plaza (1989), Gerbrandy (1999a,b), and Gerbrandy and Groeneveld (1997). These deal with actions such as α and β in our Introduction. Our examples γ and δ go beyond what is possible in the setting of these papers. But our overall setting is much more liberal setting, since it allows for all the above-mentioned types of actions. We feel it would be interesting to study further examples with an eye towards applications, but we leave this to other papers.

In our formal system, we capture only the *epistemic aspect* of these real actions, disregarding other (intentional) aspects. In particular, for simplicity reasons, we only deal with "'purely epistemic" actions; i.e., the ones that do not change the facts of the world, but affect only the agents' beliefs about the world. However, this is *not* an essential limitation, as our formal setting can be easily adapted to express fact-changing actions (see the end of section "More on Actions" and also section "Two Extensions").

On the semantical side, the main original technical contribution of our paper lies in our decision to represent not only the epistemic states, but also the epistemic actions, by Kripke structures. While for states, these structures represent in the usual way the uncertainty of each agent concerning the current state of the system, we similarly use action-structures to represent the uncertainty of each agent concerning the current action taking place. The intuition is that we are dealing with potentially "half-opaque/half-transparent" actions, about which the agents may be incompletely informed, or even completely misinformed. Besides the structure, actions have preconditions, defining their domain of applicability: not every action is possible in every state. We model the update of a state by an action as a partial *update* operation, given by a restricted product of the two structures: the uncertainties present in the given state and the given action are multiplied, while the "impossible" combinations of states and actions are eliminated (by testing the actions' preconditions on the state). The underlying intuition is that the agent's uncertainties concerning the state and the ones concerning the action are mutually independent, except for the consistency of the action with the state.

On the syntactical side, we use a mixture of dynamic and epistemic logic, with dynamic modalities associated to each action-structure, and with common-knowledge modalities for various groups of agents (in addition to the usual individual-knowledge operators). We give a complete and decidable axiomatization for this logic, and we prove various expressivity results. From a proof-theoretical point of view, the main originality of our system is the presence of our Action Rule, an inference rule capturing what might be called a notion of "epistemic (co)recursion". We understand this rule and our Knowledge-Action Axiom (a generalization of Ramsey's axiom to half-opaque actions) as expressing fundamental formal features of the interaction between action and knowledge in multi-agent systems, features that we think have not been formally expressed before.

Further Contents of This Paper

Section "A Logical Language with Epistemic Actions" gives our basic logic $\mathcal{L}([\alpha])$ of epistemic actions and knowledge. The idea is to define the logic together with the action structures which we have just looked at informally. So in $\mathcal{L}([\alpha])$ we finally will present the promised formal versions of the announcements of section "Epistemic Actions". In section "A Logic for $\mathcal{L}([\alpha])$" we present a sound

and complete axiomatization of $\mathcal{L}([\alpha])$. We add the common knowledge operators to get $\mathcal{L}([\alpha], \Box^*)$ in section "A Logic for $\mathcal{L}([\alpha], \Box^*)$". Completeness for this logic is proved in section "Completeness for $\mathcal{L}([\alpha], \Box^*)$". Two results on the expressive power are presented in section "Results on Expressive Power". An Appendix contains some technical results which, while needed for our work, seem to interrupt the flow of the paper.

A Logical Language with Epistemic Actions

Syntax

We begin with a set AtSen of atomic sentences, and we define two sets simultaneously: the language $\mathcal{L}([\alpha])$, and a set of *actions* (*over* $\mathcal{L}([\alpha])$).

$\mathcal{L}([\alpha])$ is the smallest collection which includes AtSen and which is closed under $\neg$, $\wedge$, $\Box_A$ for $A \in \mathcal{A}$, and $[\alpha]\varphi$, where α is an action over $\mathcal{L}([\alpha])$, and $\varphi \in \mathcal{L}([\alpha])$.

An *action structure* (*over* $\mathcal{L}([\alpha])$) is a pair $\langle K, \text{PRE}\rangle$, where K is a finite Kripke frame over the set $\mathcal{A}$ of agents, and PRE is a map $\text{PRE} : K \to \mathcal{L}$. We will usually write K for the action structure $\langle K, \text{PRE}\rangle$. An *action (over* $\mathcal{L}([\alpha])$*)* is a tuple $\alpha = \langle K, k, \text{PRE}\rangle$, where $\langle K, \text{PRE}\rangle$ is an action structure over $\mathcal{L}([\alpha])$, and $k \in K$. Each action α thus is a finite set with relations $\xrightarrow{D}$ for $D \in \mathcal{A}$, together with a precondition function and a specified *actual* world.

The actions themselves constitute a Kripke frame Actions in the natural way, by setting

$$\langle K, k, \text{PRE}\rangle \xrightarrow{D} \langle L, l, \text{PRE}'\rangle \quad \text{iff} \quad K = L,\ \text{PRE} = \text{PRE}',\ \text{and } k \xrightarrow{D} l \text{ in } K. \tag{38.2}$$

When $\alpha = \langle K, k, \text{PRE}\rangle$, we set $\text{PRE}(\alpha) = \text{PRE}(k)$. That is, $\text{PRE}(\alpha)$ is the precondition associated to the distinguished world of the action. For this reason, we often write $\text{PRE}(\alpha)$ instead of $\text{PRE}(k)$.

Examples 1. All of the sentences mentioned in section "Models" are sentences of $\mathcal{L}([\alpha])$, except for the ones that use $\Box^*_{\{A,B,C\}}$. This construct gives us a more expressive language, as we shall see. The structures α, τ, β, γ, γ', δ, and δ' described informally in section "Epistemic Actions" are bona fide actions. As examples of the accessibility relation on the class of actions, we have the following facts: $\alpha \xrightarrow{D} \alpha$ and $\tau \xrightarrow{D} \tau$ for all $D \in \{A, B, C\}$; $\beta \xrightarrow{B} \beta$; $\beta \xrightarrow{C} \beta$; $\beta \xrightarrow{C} \tau$; $\gamma \xrightarrow{AB} \beta$; $\gamma, \gamma' \xrightarrow{C} \gamma, \gamma'$; $\gamma' \xrightarrow{AB} \tau$; $\delta \xrightarrow{AB} \delta, \delta' \xrightarrow{AB} \delta'$, and $\delta, \delta' \xrightarrow{C} \delta, \delta'$.

Many other types of examples are possible. We can represent misleading epistemic actions, e.g. lying, or more generally acting such that some people do not suspect that your action is possible. We can also represent gratuitous suspicion ("paranoia"): maybe no "real" action has taken place, except that some people start suspecting some action (e.g., some private communication) has taken place.

Semantics

As with the syntax, we define two things simultaneously: the semantic relation $\langle W, w\rangle \models \varphi$, and a partial operation $(\langle W, w\rangle, \alpha) \mapsto \langle W, w\rangle^\alpha$. Before this, we need another definition. Given a model W and an action structure K, we define the model W^K as follows:

1. The worlds of W^K are the pairs $(w, k) \in W \times K$ such that $\langle W, w\rangle \models \text{PRE}(k)$.
2. For such pairs,

$$(w, k) \xrightarrow{A} (w', k') \quad \text{iff} \quad w \xrightarrow{A} w' \text{ and } k \xrightarrow{A} k'. \tag{38.3}$$

3. We interpret the atomic sentences by setting $v_{W^K}((w, k)) = v_W(w)$. That is, p is true at (w, k) in W^K iff p is true at w in W.

Given an action $\alpha = \langle K, k\rangle$ and a model-world pair $\langle W, w\rangle$, we say that $\langle W, w\rangle^\alpha$ is defined iff $\langle W, w\rangle \models \text{PRE}(k)$, and in that case we set $\langle W, w\rangle^\alpha = \langle W, w\rangle^{\langle K,k\rangle} = \langle W^K, (w, k)\rangle$. One can now check that the following holds for these definitions.

$$\langle W, w\rangle^\alpha \xrightarrow{A} \langle W, x\rangle^\beta \quad \text{iff} \quad \langle W, w\rangle^\alpha \text{ and } \langle W, x\rangle^\beta \text{ are defined, } w \xrightarrow{A} x \text{ in } W, \text{ and } \alpha \xrightarrow{A} \beta.$$

The semantics is given by extending the usual clauses for modal logic by one for actions:

$$\langle W, w\rangle \models [\alpha]\varphi \quad \text{iff} \quad \langle W, w\rangle^\alpha \text{ is defined implies } \langle W, w\rangle^\alpha \models \varphi.$$

As is customary, we abbreviate $\neg[\alpha]\neg\varphi$ by $\langle\alpha\rangle\varphi$. Then we have

$$\langle W, w\rangle \models \langle\alpha\rangle\varphi \quad \text{iff} \quad \langle W, w\rangle^\alpha \text{is defined and } \langle W, w\rangle^\alpha \models \varphi.$$

We also abbreviate the boolean connectives classically, and we let *true* denote some tautology such as $p \vee \neg p$.

The larger language $\mathcal{L}([\alpha], \Box^*)$ We also consider a larger language $\mathcal{L}([\alpha], \Box^*)$. This is defined by adding operators $\Box^*_{\mathcal{B}}$ for all subsets $\mathcal{B} \subseteq \mathcal{A}$. (When we do this, of course we get more actions as well.) The semantics works by taking $\Box^*_{\mathcal{B}}\varphi$ to abbreviate the infinitary conjunction

$$\bigwedge_{\langle A_1,\ldots,A_n\rangle \in \mathcal{B}^*} \Box_{A_1} \cdots \Box_{A_n}\varphi.$$

Here $\mathcal{B}^*$ is the set of all sequences from $\mathcal{B}$. This includes the empty sequence, so $\Box^*_{\mathcal{B}}\varphi$ logically implies φ.

Bisimulation Given two models, say K and L, over the same set of $\mathcal{A}$ of agents, a *bisimulation* between K and L is a relation $R \subseteq K \times L$ such that if kRl and $A \in \mathcal{A}$, then:

1. For all atomic p, $\langle K, k\rangle \models p$ iff $\langle L, l\rangle \models p$.
2. For all $k \xrightarrow{A} k'$ there is some $l \xrightarrow{A} l'$ such that $k'Rl'$.
3. For all $l \xrightarrow{A} l'$ there is some $k \xrightarrow{A} k'$ such that $k'Rl'$.

Given two model-world pairs $\langle K, k\rangle$ and $\langle L, l\rangle$, we write $\langle K, k\rangle \equiv \langle L, l\rangle$ iff there is some bisimulation R such that kRl. It is a standard fact that if $\langle K, k\rangle \equiv \langle L, l\rangle$, then the two pairs agree on all sentences of standard modal logic. In our setting, we also can speak about actions being bisimilar: we change condition (1) above to refer to say that $\text{PRE}(k) = \text{PRE}(l)$. It is easy now to check two things simultaneously: (1) bisimilar pairs agree on all sentences of $\mathcal{L}([\alpha])$; and (2) if $\langle K, k\rangle \equiv \langle L, l\rangle$ and $\alpha \equiv \beta$, then $\langle K, k\rangle^{\alpha} \equiv \langle L, l\rangle^{\beta}$. Furthermore, these results extend to $\mathcal{L}([\alpha], \Box^*)$.

Examples 2. We look back at section "Models" for some examples. We use $\cong$ to denote the relation of isomorphism on model-world pairs. It is not hard to check the following: $\langle U, u_6\rangle^{\alpha} \cong \langle V, v_6\rangle$, $\langle U, u_6\rangle^{\beta} \cong \langle X, x_6\rangle$, $\langle U, u_6\rangle^{\gamma} \cong \langle Y, y_6\rangle$, and $\langle U, u_6\rangle^{\delta} \cong \langle Z, z_6\rangle$. For example, the isomorphism which shows that $\langle U, u_6\rangle^{\delta} \cong \langle Z, z_6\rangle$ is $(u_i, o) \mapsto z_i$ for $i \neq 2, 4$, $(u_2, q) \mapsto z_2$, $(u_4, p) \mapsto z_4$, and $(u_i, r) \mapsto z'_i$ for all i.

Let α' be the action of announcing to all agents that both A and B do know whether they are dirty. Then $\langle V, v_6\rangle^{\alpha'} \cong \langle X, x_6\rangle$. Moreover, $\langle Z, v_6\rangle^{\alpha'} \equiv \langle X, x_6\rangle$. Note that in this case we only have bisimilarity. However, we know that our languages will not discriminate between bisimilar pairs, so we can regard them as the same. This models our intuition that the epistemic states at the end of Scenarios 1.5 and 4.5 should be the same.

Finally, all of the semantic facts about the various models in section "Models" now turn into precise statements. For example, $\langle U, u_6\rangle \models [\alpha] \Diamond_A \Box_B \mathsf{D}_B$. Also, $\langle U, u_6\rangle \models [\alpha][\alpha']\Box^*_{A,B,C}\Box_C \mathsf{D}_C$. This formalizes our intuition that if we start with $\langle U, u_6\rangle$, first announce that each of A and B do not know their state, then second announce that they each do know it, then at that point it will be common knowledge to all three that C knows she is dirty.

More on Actions

In this section, we have a few remarks on actions. The point here is to clarify the relation between the scenarios of section "Introduction: Example Scenarios and Their Representations" and the intuitions concerning them, and the corresponding actions of section "Epistemic Actions".

First and foremost, here are the the conceptual points involved in our formalization. The idea is that epistemic actions present a lot of uncertainty. Indeed, what might be thought of as a single action (or event) is naturally interpreted by agents

in different ways. The various agents might be unclear on what exactly happened, and again they might well have different interpretations on what is happening. Our formalization reflects this by making epistemic actions into Kripke models. So our use of possible-worlds modeling of actions is on a par with other uses of these models, and it inherits all of the features and bugs of those approaches.

Next, we want to spell out in words what our proposal amounts to. The basic problem is to decide how to represent what happens to a Kripke model W after an announcement α. (Of course, we are modeling α by an action in our formal sense.) Our solution begins by considering copies of W, one for each action token k of α in which $\text{PRE}(\alpha)$ holds. We can think of tagging the worlds of W with the worlds of α, and then we must give an account of the accessibility relation between them. The intuition is that the agents' relations to alternative worlds should be independent from their relations to other possibilities for α. So the accessibility relations of K and W should be combined *independently*. This is expressed formally in (38.3).

The auxiliary language $\hat{\mathcal{L}}$ has as atomic sentences all sentences φ of $\mathcal{L}([\alpha], \Box^*)$. It has all boolean connectives, standard modal operators $\Box_A$ for $A \in \mathcal{A}$, and also group knowledge operators $\Box^*_{\mathcal{B}}$ for $\mathcal{B} \subseteq \mathcal{A}$.

We interpret $\hat{\mathcal{L}}$ on actions using the standard clauses for the connectives and modal operators, and by interpreting the atomic sentences as follows $\langle K, k\rangle \models p$ iff $\text{PRE}(k) = p$.

Examples 3. The idea here is that the auxiliary language formalizes talk about what the different agents think is happening in our announcements. We refer back to the actions of section "Epistemic Actions". For example, $\alpha \models \Box^*_{\{A,B,C\}}\psi$. Intuitively, in α, it is common knowledge that ψ was announced. Another example: that

$$\delta \models \Box^*_{\{A,B,C\}} \Diamond_C (\psi \vee \psi_A \vee \psi_B).$$

That is, in δ, it is common knowledge that C thinks it is possible that some non-trivial announcement happened. Recall that this was one of our basic intuitions about δ, one which is not in general statable in our main language $\mathcal{L}([\alpha], \Box^*)$.

Definition. Let $\langle K, k\rangle$ be a model-world pair, and let φ be a sentence of $\hat{\mathcal{L}}$. Then χ *characterizes* $\langle K, k\rangle$ iff for all $\langle L, l\rangle$, $\langle L, l\rangle \models \chi$ iff $\langle L, l\rangle \equiv \langle K, k\rangle$.

Proposition 4. *Let $\langle K, k\rangle$ be a model-world pair with K finite. Then there is a sentence χ of $\hat{\mathcal{L}}$ which characterizes $\langle K, k\rangle$.*

Proof. By replacing $\langle K, k\rangle$ by its quotient under the largest auto-bisimulation, we may assume that if $l \neq m$, then $\langle K, l\rangle \not\equiv \langle K, m\rangle$. It is well-known that the relation of elementary equivalence in modal logic is a bisimulation on models in which each world has finitely many arrows coming in and out. It follows from this and the overall finiteness of K that we can find sentences φ_l for $l \in K$ with the property that for all l and m, $\langle K, m\rangle \models \varphi_l$ iff $m = l$. Let ψ be the following sentence

$$\psi \quad \equiv \quad \bigwedge_{l \in K, A \in \mathcal{A}} \left(\varphi_l \rightarrow \Box_A \bigvee_{l \xrightarrow{A} l'} \varphi_{l'} \wedge \bigwedge_{l \xrightarrow{A} l'} \Diamond_A \varphi_{l'} \right)$$

Going back to our original $\langle K, k \rangle$, let χ be $\varphi_k \wedge \Box^*_{\mathcal{A}} \psi$. It is easy to check that each $\langle K, l \rangle$ satisfies ψ; hence each satisfies $\Box^*_{\mathcal{A}} \psi$. Therefore $\langle K, k \rangle \models \chi$. We claim that χ characterizes $\langle K, k \rangle$. To see this, suppose that $\langle J, j \rangle \models \chi$. Consider the relation $R \subseteq K \times J$ given by

$$k' R j' \quad \text{iff} \quad \langle J, j' \rangle \models \varphi_{k'} \wedge \Box^*_{\mathcal{A}} \psi.$$

It is sufficient to see that R is a bisimulation. We'll verify half of this: suppose that $k' R j'$ and $j' \xrightarrow{A} j''$. By using ψ, we see that there is some k'' such that $k' \xrightarrow{A} k''$ and $j'' \models \varphi_{k''}$. And also, since $\models \Box^*_{\mathcal{A}} \psi \rightarrow \Box_A \Box^*_{\mathcal{A}} \psi$, we see that $\langle J, j'' \rangle \models \Box^*_{\mathcal{A}} \psi$. This completes the proof.

The connection of this result and our discussion of actions is that it is often difficult to go from an informal description of an an epistemic action to a formal one along our lines. (For example, our formulation of δ was the last of several versions.) Presumably, one way to get a formal action in our sense is to think carefully about which properties the action should have, express them in the auxiliary language, and then write a characterizing sentence such as ψ in the proof of Proposition 4. Then one can construct the finite model by standard methods. Although this would be a tedious process, it seems worthwhile to know that it is available.

Our formalization of actions reflects some choices which one might wish to modify. One of these choices is to take the range of the function PRE to be some language. Another option would be to have the range to be the *power set* of that language. This would make actions into Kripke models over the whole set of sentences. (And so what we have done is like considering modal logic with the restriction that at any world satisfies exactly one atomic sentence.) Taking this other option thus brings actions and models closer. This idea is pursued in Baltag (1999), a continuation of this work which develops a "calculus of epistemic actions." This replaces the "semantic" actions of this paper with *action expressions*. These expressions have nicer properties than the auxiliary language of this paper, but it would take us too far afield to discuss this further.

On a different matter, it makes sense to restrict attention from the full collection of actions as we have defined it to the smaller collection of *S5 actions*, where each accessibility $\xrightarrow{A}$ is an equivalence relation. This corresponds to the standard move of restricting attention to models with this property, and the reasons for doing this are similar. Intuitively, an S5 action is one in which every agent is introspective (with respect to their own suspicions about actions). Moreover, the introspection is accurate, and this fact is common knowledge.

A final modification which is quite natural is to allow actions which *change the world*. One would do this by adding to our notion of action a *sentential update u*. This would be a function defined on AtSen and written in terms of update equations

such as $u(p) := p \wedge q$; $u(q) = \mathit{false}$, etc. We are confident that our logical systems can be modified to reflect this change, and we discuss this at certain points below. We decided not to make this change mostly in order to keep the basic notions as simple as possible.

With respect to both of the changes mentioned in the last two paragraphs, it is not hard to modify our logical work to get completeness results for the new systems. We discuss all of this in section "Two Extensions".

A Logic for $\mathcal{L}([\alpha])$

In Fig. 38.2 below we present a logic for $\mathcal{L}([\alpha], \Box^*)$ which we shall study later. In this section, we shall restrict the logic to the simpler language $\mathcal{L}([\alpha])$. We do so partly to break up the study of a system with many axioms and rules, and partly to emphasize the significance of adding the infinitary operators $\Box^*_{\mathcal{B}}$ to $\mathcal{L}([\alpha])$. To carry

	Basic Axioms	
	All sentential validities	
	($\overset{\cdot}{[\alpha]}$-normality)	$\vdash [\alpha](\varphi \to \psi) \to ([\alpha]\varphi \to [\alpha]\psi)$
	($\Box_A$-normality)	$\vdash \Box_A(\varphi \to \psi) \to (\Box_A\varphi \to \Box_A\psi)$
$*$	($\sqcup^*_{\mathcal{C}}$-normality)	$\vdash \sqcup^*_{\mathcal{C}}(\varphi \to \psi) \to (\sqcup^*_{\mathcal{C}}\varphi \to \sqcup^*_{\mathcal{C}}\psi)$
	Action Axioms	
	(Atomic Permanence)	$\vdash [\alpha]p \leftrightarrow (\mathrm{PRE}(\alpha) \to p)$
	(Partial Functionality)	$\vdash [\alpha]\neg\chi \leftrightarrow (\mathrm{PRE}(\alpha) \to \neg[\alpha]\chi)$
	(Action-Knowledge)	$\vdash [\alpha]\Box_A\varphi \leftrightarrow (\mathrm{PRE}(\alpha) \to \bigwedge\{\Box_A[\beta]\varphi : \alpha \xrightarrow{A} \beta\})$
$*$	**Mix Axiom**	$\vdash \Box^*_{\mathcal{C}}\varphi \to \varphi \wedge \bigwedge\{\Box_A\Box^*_{\mathcal{C}}\varphi : A \in \mathcal{C}\}$
$*$	**Composition Axiom**	$\vdash \overset{\cdot}{[\alpha]}\overset{\cdot}{[\beta]}\varphi \leftrightarrow \overset{\cdot}{[\alpha \circ \beta]}\varphi$
	Modal Rules	
	(Modus Ponens)	From $\vdash \varphi$ and $\vdash \varphi \to \psi$, infer $\vdash \psi$
	($\overset{\cdot}{[\alpha]}$-necessitation)	From $\vdash \psi$, infer $\vdash \overset{\cdot}{[\alpha]}\psi$
	($\Box_A$-necessitation)	From $\vdash \varphi$, infer $\vdash \Box_A\varphi$
$*$	($\Box^*_{\mathcal{C}}$-necessitation)	From $\vdash \varphi$, infer $\vdash \Box^*_{\mathcal{C}}\varphi$
$*$	**Action Rule**	

Let ψ be a sentence, and let $\mathcal{C}$ be a set of agents. Let there be sentences χ_β for all β such that $\alpha \to^*_{\mathcal{C}} \beta$ (including α itself), and such that

1. $\vdash \chi_\beta \to [\beta]\psi$.
2. If $A \subset \mathcal{C}$ and $\beta \xrightarrow{A} \gamma$, then $\vdash (\chi_\beta \wedge \mathrm{PRE}(\beta)) \to \Box_A\chi_\gamma$.

From these assumptions, infer $\vdash \chi_\alpha \to [\alpha]\Box^*_{\mathcal{C}}\psi$.

Fig. 38.2 The logical system for $\mathcal{L}([\alpha], \Box^*)$. For $\mathcal{L}([\alpha])$, we drop the $*$ axioms and rules

out the restriction, we forget the axioms and rules of inference in Fig. 38.2 which are marked by a $*$. In particular $\alpha \circ \beta$ will be defined later (section "A Logic for $\mathcal{L}([\alpha], \Box^*)$").

The rules of the system are all quite standard from modal logic. The Action Axioms are the interesting new ones. In the Atomic Permanence axiom, p is an atomic sentence. The axiom then says that announcements do not change the brute fact of whether or not p holds. This axiom reflects the fact that our actions do not change any kind of local state. (We discuss an extension of our system in section "Two Extensions" where this axiom is not sound.) The Partial Functionality Axiom corresponds to the fact that the operation $\langle W, w\rangle \mapsto \langle W, w\rangle^\alpha$ is a partial function. The key axiom of the system is the Action-Knowledge Axiom, giving a criterion for knowledge after an announcement. We will check soundness of this axiom leaving checking soundness of other unstarred axioms and rules to the reader.

Proposition 1. *The Action-Knowledge Axiom*

$$[\alpha]\Box_A\varphi \leftrightarrow (\text{PRE}(\alpha) \rightarrow \bigwedge\{\Box_A[\beta]\varphi : \alpha \xrightarrow{A} \beta\})$$

is sound.

Proof. We remind the reader that the relevant definitions and notation are found in section "Semantics". Let α be the action $\langle K, k\rangle$. Fix a pair $\langle W, w\rangle$. If $\langle W, w\rangle \models \neg\text{PRE}(\alpha)$, then both sides of our biconditional hold. We therefore assume that $\langle W, w\rangle \models \text{PRE}(\alpha)$ in the rest of this proof. Assume that $\langle W, w\rangle^\alpha \models \Box_A\varphi$. Take some β such that $\alpha \xrightarrow{A} \beta$. This β is of the form $\langle K, k'\rangle$ for some k' such that $k \xrightarrow{A} k'$. Let $w \xrightarrow{A} w'$. We have two cases: $\langle W, w'\rangle \models \text{PRE}(k')$, and $\langle W, w'\rangle \models \neg\text{PRE}(k')$. In the latter case, $\langle W, w'\rangle \models [\beta]\varphi$ trivially. We'll show this in the former case, so assume $\langle W, w'\rangle \models \text{PRE}(k')$. Then (w', k') is a world of W^K, and indeed $(w, k) \xrightarrow{A} (w', k')$. Now our assumption that $\langle W, w\rangle^\alpha \models \Box_A\varphi$ implies that $\langle W^K, (w', k')\rangle \models \varphi$. This means that $\langle W, w'\rangle^\beta \models \varphi$. Hence $\langle W, w'\rangle \models [\beta]\varphi$. Since β and w' were arbitrary, $\langle W, w\rangle \models \bigwedge_\beta \Box_A[\beta]\varphi$.

The other direction is similar.

The rest of this section is devoted to the completeness result for $\mathcal{L}([\alpha])$. The reader not interested in this may omit the rest of this section, but at some points later we will refer back to the *term rewriting system* $\mathcal{R}$ which we shall describe shortly. Our completeness proof is based on a translation of $\mathcal{L}([\alpha])$ to ordinary modal logic $\mathcal{L}$. And this translation is based on a term rewriting system to be called $\mathcal{R}$.

The rewriting rules of $\mathcal{R}$ are:

$$\begin{array}{lll}
[\alpha]p & \rightsquigarrow & \text{PRE}(\alpha) \rightarrow p \\
[\alpha]\neg\psi & \rightsquigarrow & \text{PRE}(\alpha) \rightarrow \neg[\alpha]\psi \\
[\alpha](\psi \wedge \chi) & \rightsquigarrow & [\alpha]\psi \wedge [\alpha]\chi \\
[\alpha]\Box_A\psi & \rightsquigarrow & \text{PRE}(\alpha) \rightarrow \bigwedge\{\Box_A[\beta]\psi : \alpha \xrightarrow{A} \beta\}
\end{array}$$

As in all rewrite systems, we apply the rules of $\mathcal{R}$ at arbitrary subsentences of a given sentence. (For example, consider what happens with something like $[\alpha][\beta]\varphi$. We might rewrite $[\beta]\varphi$ using some rule, say to ψ. Then we might rewrite $[\alpha]\psi$ to something like $[\gamma]\psi$, etc.)

Lemma 2. *There is a relation $<$ on the sentences of $\mathcal{L}([\alpha])$ such that*

1. *$<$ is wellfounded.*
2. *For all rules $\varphi \rightsquigarrow \psi$ of $\mathcal{R}$, $\psi < \varphi$.*
3. *A sentence $\varphi \in \mathcal{L}([\alpha])$ is a normal form iff it is a modal sentence (that is, φ cannot be rewritten iff no actions occur in φ).*

This takes some work, and because the details are less important than the facts themselves, we have placed the entire matter in an Appendix to this paper. (The Appendix also discusses an extension of the rewrite system $\mathcal{R}$ to a system $\mathcal{R}^*$ for the larger language $\mathcal{L}([\alpha], \Box^*)$, so if you read it at this point you will need to keep this in mind.)

In the next result, we let $\mathcal{L}$ be ordinary modal logic over AtSen (where of course there are no actions).

Proposition 3. *There is a translation $t : \mathcal{L}([\alpha]) \to \mathcal{L}$ such that for all $\varphi \in \mathcal{L}([\alpha])$, φ is semantically equivalent to φ^t.*

Proof. Every sentence φ of $\mathcal{L}([\alpha])$ may be rewritten to a normal form. By Lemma 2, the normal forms of φ is a sentence in $\mathcal{L}$. We therefore set φ^t to be any normal form of φ, say the one obtained by carrying out leftmost reductions. The semantic equivalence follows from the fact that the rewrite rules themselves are sound, and from the fact that semantic equivalence is preserved by substitutions.

Lemma 4 (Substitution). *Let φ be any sentence, and let $\vdash \chi \leftrightarrow \chi'$. Suppose that $\varphi[p/\chi]$ comes from φ by replacing p by χ at some point, and $\varphi[p/\chi']$ comes similarly. Then $\vdash \varphi[p/\chi] \leftrightarrow \varphi[p/\chi']$.*

Proof. By induction on φ. The key point is that we have necessitation rules for each $[\alpha]$.

Theorem 5. *This logical system for $\mathcal{L}([\alpha])$ is strongly complete: $\Sigma \vdash \varphi$ iff $\Sigma \models \varphi$.*

Proof. The soundness half being easy, we only need to show that if $\Sigma \models \varphi$, then $\Sigma \vdash \varphi$. First, $\Sigma^t \models \varphi^t$. Since our system extends the standard complete proof system of modal logic, $\Sigma^t \vdash \varphi^t$. Now for each χ of $\mathcal{L}([\alpha])$, $\vdash \chi \leftrightarrow \chi^t$. (This is an easy induction on $<$ using Lemma 4.) As a result, $\Sigma \vdash \chi^t$ for all $\chi \in \Sigma$. So $\Sigma \vdash \varphi^t$. As we know $\vdash \varphi^t \leftrightarrow \varphi$. So we have our desired conclusion: $\Sigma \vdash \varphi$.

Strong completeness results of this kind may also be found in Plaza (1989) and in Gerbrandy and Groeneveld (1997). We discuss some of the history of the subject in section "Conclusions and Historical Remarks".

A Logic for $\mathcal{L}([\alpha], \Box^*)$

At this point, we turn to the completeness result for $\mathcal{L}([\alpha], \Box^*)$. It is easy to check that there is no hope of getting a *strong* completeness result (where one has arbitrary sets of hypotheses). The best one can hope for is weak completeness: $\vdash \varphi$ if and only if $\models \varphi$. Also, in contrast to our translations results for $\mathcal{L}([\alpha])$, the larger language $\mathcal{L}([\alpha], \Box^*)$ cannot be translated into $\mathcal{L}$ or even to $\mathcal{L}(\Box^*)$ (modal logic with extra modalities $\Box^*_{\mathcal{B}}$). We prove this in Theorem 2 below. So completeness results for $\mathcal{L}([\alpha], \Box^*)$ cannot simply be based on translation.

Our logical system is listed in Fig. 38.2 above. We discussed the fragment of the system which does not have the $*$ axioms and rules in section "A Logic for $\mathcal{L}([\alpha])$". The $\Box^*_{\mathcal{C}}$-normality Axiom and $\Box^*_{\mathcal{C}}$-necessitation Rule are standard, as is the Mix Axiom. We leave checking their soundness to the reader. The key features of the system are thus the Composition Axiom and the Action Rule. We begin with the Action Rule, restated below:

The Action Rule Let ψ be a sentence, and let $\mathcal{C}$ be a set of agents. Let there be sentences χ_β for all β such that $\alpha \rightarrow^*_{\mathcal{C}} \beta$ (including α itself), and such that

1. $\vdash \chi_\beta \rightarrow [\beta]\psi$.
2. If $A \in \mathcal{C}$ and $\beta \xrightarrow{A} \gamma$, then $\vdash (\chi_\beta \wedge \text{PRE}(\beta)) \rightarrow \Box_A \chi_\gamma$.

From these assumptions, infer $\vdash \chi_\alpha \rightarrow [\alpha]\Box^*_{\mathcal{C}}\psi$.

Remark. We use $\rightarrow^*_{\mathcal{C}}$ as an abbreviation for the reflexive and transitive closure of the relation $\bigcup_{A\in C} \xrightarrow{A}$. Recall that there are only finitely many β such that $\alpha \rightarrow^*_{\mathcal{C}} \beta$, since each is determined by a world of the same finite Kripke frame that determines α. So even though the Action Rule might look like it takes infinitely many premises, it really only takes finitely many.

Another point: if one so desires, the Action Rule could be replaced by a (more complicated) axiom scheme which we will not state here.

Lemma 1. *$\langle W, w\rangle \models \langle\alpha\rangle \Diamond^*_{\mathcal{C}} \varphi$ iff there is a sequence of worlds from W*

$$w \;=\; w_0 \quad \rightarrow_{A_1} \quad w_1 \quad \rightarrow_{A_2} \quad \cdots \quad \rightarrow_{A_{k-1}} \quad w_{k-1} \quad \rightarrow_{A_k} \quad w_k$$

where $k \geq 0$, and also a sequence of actions of the same length k,

$$\alpha \;=\; \alpha_0 \quad \rightarrow_{A_1} \quad \alpha_1 \quad \rightarrow_{A_2} \quad \cdots \quad \rightarrow_{A_{k-1}} \quad \alpha_{k-1} \quad \rightarrow_{A_k} \quad \alpha_k$$

such that $A_i \in \mathcal{C}$ and $\langle W, w_i\rangle \models \text{PRE}(\alpha_i)$ for all $0 \leq i \leq k$, and $\langle W, w_k\rangle \models \langle\alpha_k\rangle\varphi$.

Remark. The case $k = 0$ just says that $\langle W, w\rangle \models \langle\alpha\rangle \Diamond^*_{\mathcal{C}} \varphi$ is implied by $\langle W, w\rangle \models \langle\alpha\rangle\varphi$.

Proof. $\langle W, w\rangle \models \langle\alpha\rangle \Diamond^*_{\mathcal{C}} \varphi$ iff $\langle W, w\rangle \models \text{PRE}(\alpha)$ and $\langle W^\alpha, (w, \alpha)\rangle \models \Diamond^*_{\mathcal{C}}\varphi$; iff $\langle W, w\rangle \models \text{PRE}(\alpha)$ and there is a sequence in W^α,

$$(w,\alpha) \;=\; v_0 \quad \to_{A_1} \quad v_1 \quad \to_{A_2} \quad \cdots \quad \to_{A_{k-1}} \quad v_{k-1} \quad \to_{A_k} \quad v_k$$

where $k \geq 0$ such that $A_i \in \mathcal{C}$ and $\langle W^\alpha, v_k\rangle \models \varphi$. Now suppose such sequences exist in W^α. Then we get a sequence of worlds w_i in W and actions α_i such that $v_i = (w_i, \alpha_i)$ and $\langle W, w_i\rangle \models \text{PRE}(\alpha_i)$. The condition that $\langle W^\alpha, v_k\rangle \models \varphi$ translates to $\langle W, w_k\rangle \models \langle\alpha_k\rangle\varphi$. Conversely, if we have a sequence in W with these properties, we get one in W^α by taking $v_i = (w_i, \alpha_i)$.

Proposition 2. *The Action Rule is sound.*

Proof. Assume that $\langle W, w\rangle \models \chi_\alpha$ but also $\langle W, w\rangle \models \langle\alpha\rangle \Diamond_{\mathcal{C}}^* \neg\psi$. According to Lemma 1, there is a labeled sequence of worlds from W

$$w \;=\; w_0 \quad \to_{A_1} \quad w_1 \quad \to_{A_2} \quad \cdots \quad \to_{A_{k-1}} \quad w_{k-1} \quad \to_{A_k} \quad w_k$$

where $k \geq 0$ and each $A_i \in \mathcal{C}$, and also a sequence of actions of length k, with the same labels,

$$\alpha \;=\; \alpha_0 \quad \to_{A_1} \quad \alpha_1 \quad \to_{A_2} \quad \cdots \quad \to_{A_{k-1}} \quad \alpha_{k-1} \quad \to_{A_k} \quad \alpha_k$$

such that $\langle W, w_i\rangle \models \text{PRE}(\alpha_i)$ for all $0 < i \leq k$, and $\langle W, w_k\rangle \models \langle\alpha_k\rangle\neg\psi$. If $k = 0$, we have $\langle W, w\rangle \models \langle\alpha\rangle\neg\psi$. But since $\vdash \chi_\alpha \to [\alpha]\psi$, we have $\langle W, w\rangle \models [\alpha]\psi$. This is a contradiction.

Now we argue the case $k > 0$. We show by induction on $1 \leq i \leq k$ that $\langle W, w_i\rangle \models \chi_{\alpha_i} \wedge [\alpha_i]\psi$. In particular, $\langle W, w_k\rangle \models [\alpha_k]\psi$. This is a contradiction.

We close with a discussion of the Composition Rule, beginning with a general definition.

Definition. Let $\alpha = \langle K, k\rangle$ and $\beta = \langle L, l\rangle$ be actions. Then the *action composition* $\alpha \circ \beta$ is the action defined as follows. Consider the product set $K \times L$. We turn this into a Kripke frame using the restriction of the product arrows. We get an action structure by setting

$$\text{PRE}((k', l')) \quad = \quad \text{PRE}(k') \wedge [\langle K, k'\rangle]\text{PRE}(l').$$

Finally, we set $\alpha \circ \beta = \langle K \times L, (k, l)\rangle$.

Proposition 3. *Concerning the composition operation:*

1. $(W^\alpha)^\beta \cong W^{\alpha\circ\beta}$ *via the restriction of* $((w, k'), l') \mapsto (w, (k', l'))$ *to* $(W^\alpha)^\beta$.
2. *The Composition Axiom is sound:* $[\alpha][\beta]\varphi \leftrightarrow [\alpha \circ \beta]\varphi$.
3. $\alpha \circ (\beta \circ \gamma) \cong (\alpha \circ \beta) \circ \gamma$.
4. $\alpha \circ \tau \cong \alpha \cong \tau \circ \alpha$, *where the null action* τ *is from section "Epistemic Actions".*

Proof. Let $\alpha = \langle K, k\rangle$ and $\beta = \langle L, l\rangle$. For (1), note that the worlds of $(W^\alpha)^\beta$ are of the form $((w, k'), l')$, where $(w, k') \in W^\alpha$ and $\langle W^\alpha, (w, k')\rangle \models \text{PRE}(l')$. For such

$((w,k'),l')$, $\langle W,w\rangle \models \text{PRE}(k')$ and $\langle W,w\rangle \models [\langle K,k'\rangle]\text{PRE}(l')$. That is, $(w,(k',l')) \in W^{\alpha\circ\beta}$. The converse is similar, and the rest of the isomorphism properties are easy.

Part (2) follows from (1). We use the obvious isomorphism $((k,l),m) \mapsto (k,(l,m))$ in part (3). We use the Composition and $[\alpha]$-necessitation axioms to show that this isomorphism preserves the PRE function up to logical equivalence. Part (4) is easy, using the fact that $\models [\tau]\varphi \leftrightarrow \varphi$.

Extending the rewriting system $\mathcal{R}$ to $\mathcal{L}([\alpha],\Box^*)$. We consider $\mathcal{L}([\alpha],\Box^*)$. The rewriting system $\mathcal{R}$ extends naturally to this larger language, taking new symbols for the operators $\Box^*_{\mathcal{B}}$. We also add a rule corresponding to the Composition Axiom: $[\alpha][\beta]\varphi \rightsquigarrow [\alpha\circ\beta]\varphi$. We call this rewriting system $\mathcal{R}^*$.

Lemma 4. *There is a relation $<$ on the sentences and actions of $\mathcal{L}([\alpha],\Box^*)$ such that*

1. *$<$ is wellfounded.*
2. *For all rules $\varphi \rightsquigarrow \psi$ of $\mathcal{R}^*$, $\psi < \varphi$.*
3. *If ψ is a proper subsentence of φ, then $\psi < \varphi$.*
4. *A sentence $\varphi \in \mathcal{L}([\alpha],\Box^*)$ is a normal form iff it is built from atomic sentences using $\neg$, $\wedge$, $\Box_A$, and $\Box^*_{\mathcal{B}}$, or if it is of the form $[\alpha]\Box^*_{\mathcal{B}}\psi$, where α is an action in normal form, and ψ too is in normal form.*
5. *An action α is a normal form if whenever $\alpha \xrightarrow{*} \beta$, $\text{PRE}(\beta)$ is a normal form sentence.*
6. *If $\alpha \xrightarrow{*} \beta$, then $[\alpha]\Box^*_{\mathcal{C}}\psi > [\beta]\psi$.*
7. *$nf(\varphi) \leq \varphi$.*

Once again, the details are in the Appendix.

In section "A Logic for $\mathcal{L}([\alpha])$", we saw a translation t from $\mathcal{L}([\alpha])$ to $\mathcal{L}$ can be extended to a translation from $\mathcal{L}([\alpha],\Box^*)$ to the infinitary language $\mathcal{L}_\infty$, where we have countable conjunctions and disjunctions. This extension is defined using Part (4) of Lemma 4. The additional clauses in the definition of t are

$$\begin{aligned}(\Box^*_{\mathcal{B}}\varphi)^t &= \textstyle\bigwedge_{\langle A_1,\ldots,A_n\rangle\in\mathcal{B}^*}(\Box_{A_1}\cdots\Box_{A_n}\varphi)^t\\ ([\alpha]\Box^*_{\mathcal{B}}\psi)^t &= \textstyle\bigwedge_{\langle A_1,\ldots,A_n\rangle\in\mathcal{B}^*}([\alpha]\Box_{A_1}\cdots\Box_{A_n}\psi)^t\end{aligned}$$

In this way, we see that $\mathcal{L}([\alpha],\Box^*)$ may be regarded as a fragment of infinitary modal logic.

Remark. It is possible to drop the Composition Axiom in favor of a more involved version of the Action Rule. The point is the Composition Axiom simplifies the normal forms of the $\mathcal{L}([\alpha],\Box^*)$: Without the Composition Axiom, the normal forms of sentences of $\mathcal{L}([\alpha],\Box^*)$ would be of the form $[\alpha_1][\alpha_2]\cdots[\alpha_r]\psi$, where each α_i is a normal form action and ψ is a normal form sentence. The Composition Axiom insures that the normal forms are of the form $[\alpha]\psi$. So if we were to drop the Composition Axiom, we would need a formulation of the Action Rule which involved *sequences* of actions. It is not terribly difficult to formulate such a rule,

and completeness can be obtained by an elaboration of the work which we shall do. We did not present this work, mostly because adding the Composition Axiom leads to shorter proofs.

This completes the discussion of the axioms and rules of our logical system for $\mathcal{L}([\alpha], \Box^*)$.

Completeness for $\mathcal{L}([\alpha], \Box^*)$

In this section, we prove the completeness of the logical system for $\mathcal{L}([\alpha], \Box^*)$. Section "Some Syntactic Results" has some technical results which culminate in the Substitution Lemma 3. This is used in some of our work on normal forms in the Appendix, and that work figures in the completeness theorem of section "Completeness".

Some Syntactic Results

Lemma 1. *For all $A \in \mathcal{C}$ and all β such that $\alpha \rightarrow_A \beta$,*

1. $\vdash [\alpha]\Box^*_{\mathcal{C}}\psi \rightarrow [\alpha]\psi$.
2. $\vdash [\alpha]\Box^*_{\mathcal{C}}\psi \wedge \text{PRE}(\alpha) \rightarrow \Box_A[\beta]\Box^*_{\mathcal{C}}\psi$.

Proof. Part (1) follows easily from the Mix Axiom and modal reasoning. For part (2), we start with a consequence of the Mix Axiom: $\vdash \Box^*_{\mathcal{C}}\psi \rightarrow \Box_A\Box^*_{\mathcal{C}}\psi$. Then by modal reasoning, $\vdash [\alpha]\Box^*_{\mathcal{C}}\psi \rightarrow [\alpha]\Box_A\Box^*_{\mathcal{C}}\psi$. By the Action-Knowledge Axiom, we have $\vdash [\alpha]\Box^*_{\mathcal{C}}\psi \wedge \text{PRE}(\alpha) \rightarrow \Box_A[\beta]\Box^*_{\mathcal{C}}\psi$.

Definition. Let α and α' be actions. We write $\vdash \alpha \leftrightarrow \alpha'$ if α and α' are based on the same Kripke frame W and the same world w, and if for all $v \in W$, $\vdash \text{PRE}(v) \leftrightarrow \text{PRE}'(v)$, where PRE is the announcement function for α, and PRE' for α'.

We note the following bisimulation-like properties:

1. If $\vdash \alpha \leftrightarrow \alpha'$, then also $\vdash \text{PRE}(\alpha) \leftrightarrow \text{PRE}(\alpha')$.
2. Whenever β' is such that $\alpha' \rightarrow^*_{\mathcal{C}} \beta'$, then there is some β such that $\vdash \beta \leftrightarrow \beta'$ and $\alpha \rightarrow^*_{\mathcal{C}} \beta$.

These follow easily from the way we defined PRE on actions in terms of functions on frames.

Lemma 2. *If $\vdash \alpha \leftrightarrow \alpha'$, then for all ψ, $\vdash [\alpha]\psi \leftrightarrow [\alpha']\psi$.*

Proof. By induction on ψ. For ψ atomic, our result is easy. The induction steps for $\neg$ and $\wedge$ are trivial. The step for $\Box_A$ is not hard, and so we omit it. Assuming the result for ψ gives the result for $[\chi]\psi$, using the Composition Axiom and the induction hypothesis. This leaves the step for sentences of the form $\Box^*_{\mathcal{B}}\psi$, assuming

the result for ψ. We use the Action Rule to show that $\vdash [\alpha]\Box_C^*\psi \rightarrow [\alpha']\Box_C^*\psi$. For each β', we let $\chi_{\beta'}$ be $[\beta]\Box_C^*\psi$, where β is such that $\vdash \beta \leftrightarrow \beta'$. We need to show that for all relevant β' and γ',

(a) $\vdash [\beta]\Box_C^*\psi \rightarrow [\beta']\psi$; and
(b) If $\beta' \xrightarrow{A} \gamma'$, then $\vdash [\beta]\Box_C^*\psi \wedge \text{PRE}(\beta') \rightarrow \Box_A[\gamma]\Box_C^*\psi$.

For (a), we know from Lemma 1, part (1) that $\vdash [\beta]\Box_C^*\psi \rightarrow [\beta]\psi$. By induction hypothesis on ψ, $\vdash [\beta]\psi \leftrightarrow [\beta']\psi$. And this implies (a). For (b), Lemma 1, part (2) tells us that under the assumptions,

$$\vdash [\beta]\Box_C^*\psi \wedge \text{PRE}(\beta) \rightarrow \Box_A[\gamma]\Box_C^*\psi.$$

As we know, $\vdash \text{PRE}(\beta) \leftrightarrow \text{PRE}(\beta')$. This implies (b).

This completes the induction on ψ.

Lemma 3 (Substitution). *Let t be a sentence or action of $\mathcal{L}([\alpha], \Box^*)$, and let $\vdash \chi \leftrightarrow \chi'$. Suppose that $t[p/\chi]$ comes from t by replacing p by χ at some point, and $t[p/\chi']$ comes similarly. Then $\vdash t[p/\chi] \leftrightarrow t[p/\chi']$.*

Proof. By induction on t, using Lemma 2.

Lemma 4. *For every sentence $\varphi \in \mathcal{L}([\alpha], \Box^*)$ there is some normal form $nf(\varphi) \leq \varphi$ such that $\vdash \varphi \leftrightarrow nf(\varphi)$.*

Proof. Given φ, there is a finite sequence $\varphi_0 \rightsquigarrow \cdots \rightsquigarrow \varphi_n = \varphi'$ such that $\varphi_0 = \varphi$, and φ_n is in normal form. This is a consequence of the fact that $<$ is wellfounded and the rules of the system are reducing. By Lemma 3, we see that for all i, $\vdash \varphi_i \leftrightarrow \varphi_i'$.

Completeness

The proof of completeness and decidability is based on the filtration argument for completeness of PDL due to Kozen and Parikh (1981). We show that every consistent φ has a finite model, and that the size of the model is recursive in φ. We shall need to use some results concerning the rewriting system $\mathcal{R}^*$ from section "A Logic for $\mathcal{L}([\alpha], \Box^*)$".

Definition. Let $s(\varphi)$ be the set of subsentences of φ, including φ itself. This includes all sentences occurring in actions which occur in φ and their subsentences. For future use, we note that

$$s([\alpha]\Box_C^*\varphi) = \{[\alpha]\Box_C^*\varphi, \Box_C^*\varphi\} \cup s(\varphi) \cup \bigcup\{s(\text{PRE}(\beta)) : \alpha \rightarrow_C^* \beta\} \quad (38.4)$$

We define a function $f : \mathcal{L}([\alpha], \Box^*) \rightarrow \mathcal{P}(\mathcal{L}([\alpha], \Box^*))$ by recursion on the wellfounded relation $<$ as follows: For *normal forms*, f works as follows:

$$
\begin{array}{lcl}
f(p) & = & \{p\} \\
f(\neg\varphi) & = & f(\varphi) \cup \{\neg\varphi\} \\
f(\varphi \wedge \psi) & = & f(\varphi) \cup f(\psi) \cup \{\varphi \wedge \psi\} \\
f(\Box_A \varphi) & = & f(\varphi) \cup \{\Box_A \varphi\} \\
f(\Box^*_{\mathcal{B}} \varphi) & = & f(\varphi) \cup \{\Box^*_{\mathcal{B}} \varphi\} \cup \{\Box_A \Box^*_{\mathcal{B}} \varphi : A \in \mathcal{B}\} \\
f([\alpha]\Box^*_{\mathcal{C}} \varphi) & = & \{\Box_A [\beta] \Box^*_{\mathcal{C}} \varphi \ : \ \alpha \rightarrow^*_{\mathcal{C}} \beta \ \& \ A \in \mathcal{C}\} \\
 & & \cup \{[\beta]\Box^*_{\mathcal{C}} \varphi \ : \ \alpha \rightarrow^*_{\mathcal{C}} \beta \ \& \ A \in \mathcal{C}\} \\
 & & \cup \bigcup \{f(\chi) \ : \ (\exists \beta)\, \alpha \rightarrow^*_{\mathcal{C}} \beta \ \& \ \chi \in s(\mathrm{PRE}(\beta))\} \\
 & & \cup f(\Box^*_{\mathcal{C}} \varphi) \\
 & & \cup \bigcup \{f([\beta]\varphi) \ : \ \alpha \rightarrow^*_{\mathcal{C}} \beta\}
\end{array}
$$

For φ not in normal form, let $f(\varphi) = f(nf\,(\varphi))$. (Note that we need to define f on sentences which are not normal forms, because $f([\beta]\psi)$ figures in $f([\alpha]\Box^*_{\mathcal{C}}\varphi)$. Also, the definition makes sense because the calls to f on the right-hand sides are all $<$ the arguments on the left-hand sides, and since $nf\,(\varphi) \leq \varphi$ for all φ; see Lemma 4.)

Lemma 5. *For all φ:*

1. *$f(\varphi)$ is a finite set of normal form sentences.*
2. *$nf\,(\varphi) \in f(\varphi)$.*
3. *If $\psi \in f(\varphi)$, then $f(\psi) \subseteq f(\varphi)$.*
4. *If $\psi \in f(\varphi)$, then $s(\psi) \subseteq f(\varphi)$.*
5. *If $[\gamma]\Box^*_{\mathcal{C}}\chi \in f(\varphi)$, $\gamma \rightarrow^*_{\mathcal{C}} \delta$, and $A \in \mathcal{C}$, then $f(\varphi)$ also contains $\Box_A[\delta]\Box^*_{\mathcal{C}}\chi$, $[\delta]\Box^*_{\mathcal{C}}\chi$, $\mathrm{PRE}(\delta)$, and $nf\,([\delta]\chi)$.*

Proof. All of the parts are by induction on φ in the well-order $<$. For part (1), note that if $[\alpha]\Box^*_{\mathcal{C}}\psi$ is a normal form, then each sentence $\Box_A[\beta]\Box^*_{\mathcal{C}}\psi$ and all subsentences of this sentence are normal forms. For part (2), note that when φ is a normal form, $\varphi \in f(\varphi)$.

In part (3), we only need to consider φ in normal form. The result is immediate when φ is an atomic sentence p. The induction steps for $\neg$, $\wedge$, and $\Box_A$ are easy. For $\Box^*_{\mathcal{B}}\varphi$, note that since $\varphi < \Box^*_{\mathcal{B}}\varphi$, our induction hypothesis implies the result for φ; we verify it for $\Box^*_{\mathcal{B}}\varphi$. The only interesting case is when ψ is $\Box_A\Box^*_{\mathcal{B}}\varphi$ for some $A \in \mathcal{B}$. And in this case

$$f(\psi) \quad = \quad f(\Box^*_{\mathcal{B}}\varphi) \cup \{\Box_A\Box^*_{\mathcal{B}}\varphi\} \quad \subseteq \quad f(\Box^*_{\mathcal{B}}\varphi).$$

To complete part (3), we consider $[\alpha]\Box^*_{\mathcal{C}}\varphi$. If there is some $\chi < [\alpha]\Box^*_{\mathcal{C}}\varphi$ such that $\psi \in f(\chi)$ and $f(\chi) \subseteq f([\alpha]\Box^*_{\mathcal{C}}\varphi)$, then we are easily done by the induction hypothesis. This covers all of the cases except for $\psi = [\beta]\Box^*_{\mathcal{C}}\varphi$ and $\psi = \Box_A[\beta]\Box^*_{\mathcal{C}}\varphi$. For the first of these, we use the transitivity of $\rightarrow^*_{\mathcal{C}}$ to check that $f([\beta]\Box^*_{\mathcal{C}}\varphi) \subseteq f([\alpha]\Box^*_{\mathcal{C}}\varphi)$. And now the second case follows:

$$f(\Box_A[\beta]\Box^*_{\mathcal{C}}\varphi) \quad = \quad f([\beta]\Box^*_{\mathcal{C}}\varphi) \cup \{\Box_A[\beta]\Box^*_{\mathcal{C}}\varphi\} \quad \subseteq \quad f([\alpha]\Box^*_{\mathcal{C}}\varphi).$$

Part (4) is similar to part (3), using equation (38.4) at the beginning of this subsection.

For part (5), assume that $[\gamma]\Box^*_{\mathcal{C}}\chi \in f(\varphi)$. By part (1), $[\gamma]\Box^*_{\mathcal{C}}\chi$ is a normal form. We show that $\Box_A[\delta]\Box^*_{\mathcal{C}}\chi$, $[\delta]\Box^*_{\mathcal{C}}\chi$, PRE($\delta$), and $nf([\delta]\chi)$ all belong to $f([\gamma]\Box^*_{\mathcal{C}}\chi)$, and then use part (3). The first two of these sentences are immediate by the definition of f; the third one follows from part (4); and the last comes from part (2) since $nf([\delta]\chi) \in f([\delta]\chi) \subseteq f([\gamma]\Box^*_{\mathcal{C}}\chi$.

The set $\Delta = \Delta(\varphi)$ Fix a sentence φ. We set $\Delta = f(\varphi)$ (i.e., we drop φ from the notation). This set Δ is the version for our logic of the Fischer-Ladner closure of φ. Let $\Delta = \{\psi_1, \ldots, \psi_n\}$. Given a maximal consistent set U of $\mathcal{L}([\alpha], \Box^*)$, let

$$[\![U]\!] \quad = \quad \underline{+}\psi_1 \wedge \cdots \wedge \underline{+}\psi_n,$$

where the signs are taken in accordance with membership in U. That is, if $\psi_i \in U$, then ψ is a conjunct of $[\![U]\!]$; but if $\psi_i \notin U$, then $\neg\psi_i$ is a conjunct.

Two (standard) observations are in order. Notice that if $[\![U]\!] \neq [\![V]\!]$, then $[\![U]\!] \wedge [\![V]\!]$ is inconsistent. Also, for all $\psi \in \Delta$,

$$\vdash \psi \leftrightarrow \bigvee\{[\![W]\!] : W \text{ is maximal consistent and } \psi \in W\}. \tag{38.5}$$

and

$$\vdash \neg\psi \leftrightarrow \bigvee\{[\![W]\!] : W \text{ is maximal consistent and } \neg\psi \in W\}. \tag{38.6}$$

(The reason is that ψ is equivalent to the disjunction of *all* complete conjunctions which contain it. However, some of those complete conjunctions are inconsistent and these can be dropped from the big disjunction. The others are consistent and hence can be extended to maximal consistent sets.)

Definition. The *filtration* $\mathcal{F}$ is the model whose worlds are the equivalence classes $[U]$, where U is a maximal consistent set in the logic for $\mathcal{L}([\alpha], \Box^*)$, and the equivalence relation is $U \equiv V$ iff $[\![U]\!] = [\![V]\!]$ (iff $U \cap \Delta = V \cap \Delta$). We set $\langle \mathcal{F}, [U] \rangle \models p$ iff $p \in U \cap \Delta$. Furthermore,

$$[U] \xrightarrow{A} [V] \text{ in } \mathcal{F} \quad \text{iff} \quad \text{whenever } \Box_A\psi \in U \cap \Delta, \text{ then also } \psi \in V\,. \tag{38.7}$$

This condition is independent of the choice of representatives: we use part (4) of Lemma 5 to see that if $\Box_A\chi \in \Delta$, then also $\chi \in \Delta$.

A *good path from* $[V_0]$ *for* $\langle\alpha\rangle\, \Diamond^*_{\mathcal{C}}\, \psi$ is a path in $\mathcal{F}$

$$[V_0] \quad \rightarrow_{A_1} \quad [V_1] \quad \rightarrow_{A_2} \quad \cdots \quad \rightarrow_{A_{k-1}} \quad [V_{k-1}] \quad \rightarrow_{A_k} \quad [V_k]$$

such that $k \geq 0$, each $A_i \in \mathcal{C}$, and such that there exist actions

$$\alpha = \alpha_0 \quad \rightarrow_{A_1} \quad \alpha_1 \quad \rightarrow_{A_2} \quad \cdots \quad \rightarrow_{A_{k-1}} \quad \alpha_{k-1} \quad \rightarrow_{A_k} \quad \alpha_k$$

such that $\text{PRE}(\alpha_i) \in V_i$ for all $0 \leq i \leq k$, and $\langle \alpha_k \rangle \psi \in V_k$.

The idea behind a good path comes from considering Lemma 1 in $\mathcal{F}$. Of course, the special case of that result would require that $\langle \mathcal{F}, [V_i] \rangle \models \text{PRE}(\alpha_i)$ rather than $\text{PRE}(\alpha_i) \in V_i$, and similarly for $\langle \alpha_k \rangle \psi$ and V_k. The exact formulation above was made in order that the Truth Lemma will go through for sentences of the form $\langle \alpha \rangle \Diamond_{\mathcal{C}}^* \psi$ (see the final paragraphs of the proof of Lemma 8).

Lemma 6. *Let* $[\alpha] \Box_{\mathcal{C}}^* \psi \in \Delta$. *If there is a good path from* $[V_0]$ *for* $\langle \alpha \rangle \Diamond_{\mathcal{C}}^* \neg \psi$, *then* $\langle \alpha \rangle \Diamond_{\mathcal{C}}^* \neg \psi \in V_0$.

Proof. By induction on the length k of the path. If $k = 0$, then $\langle \alpha \rangle \neg \psi \in V_0$. If $\langle \alpha \rangle \Diamond_{\mathcal{C}}^* \neg \psi \notin V_0$, then $[\alpha] \Box_{\mathcal{C}}^* \psi \in V_0$. By Lemma 1, part (1), we have $[\alpha] \psi \in V_0$. This is a contradiction.

Assume the result for k, and suppose that there is a good path from $[V_0]$ for $\langle \alpha \rangle \Diamond_{\mathcal{C}}^* \neg \psi$ of length $k + 1$. Then there is a good path of length k from $[V_1]$ for $\langle \alpha_1 \rangle \Diamond_{\mathcal{C}}^* \neg \psi$. Also, $[\alpha_1] \Box_{\mathcal{C}}^* \psi \in \Delta$, by Lemma 5, part (5). By induction hypothesis, $\langle \alpha_1 \rangle \Diamond_{\mathcal{C}}^* \neg \psi \in V_1$.

If $\langle \alpha \rangle \Diamond_{\mathcal{C}}^* \neg \psi \notin V_0$, then $[\alpha] \Box_{\mathcal{C}}^* \psi \in V_0$. By Lemma 1, part (2), V_0 contains $[\alpha] \Box_{\mathcal{C}}^* \psi \wedge \text{PRE}(\alpha) \rightarrow \Box_A [\alpha_1] \Box_{\mathcal{C}}^* \psi$. So V_0 contains $\Box_A [\alpha_1] \Box_{\mathcal{C}}^* \psi$. Again, this sentence belongs to Δ by Lemma 5, part (5). Now by definition of $\stackrel{A}{\rightarrow}$ in $\mathcal{F}$, we see that $[\alpha_1] \Box_{\mathcal{C}}^* \psi \in V_1$. This is a contradiction.

Lemma 7. *If* $[\![V_0]\!] \wedge \langle \alpha \rangle \Diamond_{\mathcal{C}}^* \psi$ *is consistent, then there is a good path from* $[V_0]$ *for* $\langle \alpha \rangle \Diamond_{\mathcal{C}}^* \psi$.

Proof. For each β such that $\alpha \rightarrow_{\mathcal{C}}^* \beta$, let S_β be the (finite) set of all $[W] \in \mathcal{F}$ such that there is *no* good path from $[W]$ for $\langle \beta \rangle \Diamond_{\mathcal{C}}^* \psi$. We need to see that $[V_0] \notin S_\alpha$; suppose toward a contradiction that $[V_0] \in S_\alpha$. Let

$$\chi_\beta \quad = \quad \bigvee \{ [\![W]\!] : W \in S_\beta \}.$$

Note that $\neg \chi_\beta$ is logically equivalent to $\bigvee \{ [\![W']\!] : [W'] \in \mathcal{F} \text{and} W' \notin S_\beta \}$. Since we assumed $V_0 \in S_\alpha$, we have $\vdash [\![V_0]\!] \rightarrow \chi_\alpha$.

We first claim that $\chi_\beta \wedge \langle \beta \rangle \psi$ is inconsistent. Otherwise, there would be $[W] \in S_\beta$ such that $\chi_\beta \wedge \langle \beta \rangle \psi \in W$. Note that by the Partial Functionality Axiom, $\vdash \langle \beta \rangle \psi \rightarrow \text{PRE}(\beta)$. But then the one-point path $[W]$ is a good path from $[W]$ for $\langle \beta \rangle \Diamond_{\mathcal{C}}^* \psi$. Thus $[W] \notin S_\beta$, and this is a contradiction. So indeed, $\chi_\beta \wedge \langle \beta \rangle \psi$ is inconsistent. Therefore, $\vdash \chi_\beta \rightarrow [\beta] \neg \psi$.

We will need the following standard claim, an argument for which can be found in Kozen and Parikh (1981). We will also use this claim in the proof of Lemma 8.

Claim. If $[\![U]\!] \wedge \Diamond_A [\![V]\!]$ is consistent, then $[U] \rightarrow_A [V]$.

Proof of Claim. Assume $\Box_A\psi \in U \cap \Delta$. If $\psi \notin V$, then $\neg\psi \in V$, so since $\psi \in \Delta$, $\vdash [\![V]\!] \rightarrow \neg\psi$. Thus, $\vdash \Diamond_A[\![V]\!] \rightarrow \Diamond_A\neg\psi$, and so $\vdash [\![U]\!] \wedge \Diamond_A[\![V]\!] \rightarrow \Box_A\psi \wedge \Diamond_A\neg\psi$, whence $[\![U]\!] \wedge \Diamond_A[\![V]\!]$ is inconsistent. This contradiction establishes the claim.

We next show that for all $A \in \mathcal{C}$ and all β such that $\beta \rightarrow_A \gamma$, $\chi_\beta \wedge \text{PRE}(\beta) \wedge \Diamond_A\neg\chi_\gamma$ is inconsistent. Otherwise, there would be $[W] \in S_\beta$ with χ_β, $\text{PRE}(\beta)$, and $\Diamond_A\neg\chi_\gamma$ in it. Then $\bigvee\{\Diamond_A[\![W']\!] : W' \notin S_\gamma\}$, being equivalent to $\Diamond_A\neg\chi_\beta$, would belong to W. It follows that $\Diamond_A[\![W']\!] \in W$ for some $W' \notin S_\gamma$. By the claim, $[W] \rightarrow_A [W']$. Since $[W'] \notin S_\gamma$, there is a good path from $[W']$ for $\langle\gamma\rangle\Diamond_{\mathcal{C}}^*\psi$. But since $\beta \rightarrow_A \gamma$ and W contains $\text{PRE}(\beta)$, we also have a good path from $[W]$ for $\langle\beta\rangle\Diamond_{\mathcal{C}}^*\psi$. This again contradicts $[W] \in S_\beta$. As a result, for all relevant A, β, and γ, $\vdash \chi_\beta \wedge \text{PRE}(\beta) \rightarrow \Box_A\chi_\gamma$.

By the Action Rule, $\vdash \chi_\alpha \rightarrow [\alpha]\Box_{\mathcal{C}}^*\neg\psi$. Now $\vdash [\![V_0]\!] \rightarrow \chi_\alpha$. So $\vdash [\![V_0]\!] \rightarrow [\alpha]\Box_{\mathcal{C}}^*\neg\psi$. This contradicts the assumption with which we began this proof.

Lemma 8 (Truth Lemma). *Consider a sentence φ, and also the set $\Delta = f(\varphi)$. For all $\chi \in \Delta$ and $[U] \in \mathcal{F}$: $\chi \in U$ iff $\langle\mathcal{F}, [U]\rangle \models \chi$.*

Proof. We argue by induction on the wellfounded $<$ that if $\chi \in \Delta$, then: $\chi \in U$ iff $\langle\mathcal{F}, [U]\rangle \models \chi$. The case of χ atomic is trivial. Now assume this Truth Lemma for sentences $< \chi$. Note that by soundness, we may assume that χ is in normal form. We argue by cases on χ.

The cases that χ is either a negation or conjunction are trivial.

Suppose next that $\chi \equiv \Box_A\psi$. Suppose $\Box_A\psi \in U$; we show $\langle\mathcal{F}, [U]\rangle \models \Box_A\psi$. Let $[V]$ be such that $[U] \stackrel{A}{\rightarrow} [V]$. Then by definition of $\stackrel{A}{\rightarrow}$, $\psi \in V$. The induction hypothesis applies to ψ, since $\psi < \Box_A\psi$, and since $\psi \in \Delta$ by Lemma 5, part (4). So by induction hypothesis, $\langle\mathcal{F}, [V]\rangle \models \psi$. This gives half of our equivalence. Conversely, suppose that $\langle\mathcal{F}, [U]\rangle \models \Box_A\psi$. Suppose towards a contradiction that $\Diamond_A\neg\psi \in U$. So $[\![U]\!] \wedge \Diamond_A\neg\psi$ is consistent. We use Eq. (38.6) and the fact that $\Diamond_A$ distributes over disjunctions to see that $[\![U]\!] \wedge \Diamond_A\neg\psi$ is logically equivalent to $\bigvee([\![U]\!] \wedge \Diamond_A[\![V]\!])$, where the disjunction is taken over all V which contain $\neg\psi$. Since $[\![U]\!] \wedge \Diamond_A\neg\psi$ is consistent, one of the disjuncts $[\![U]\!] \wedge \Diamond_A[\![V]\!]$ must be consistent. The induction hypothesis again applies, and we use it to see that $\langle\mathcal{F}, [V]\rangle \models \neg\psi$. By the claim in the proof of Lemma 7, $[U] \stackrel{A}{\rightarrow} [V]$. We conclude that $\langle\mathcal{F}, [U]\rangle \models \Diamond_A\neg\psi$, and this is a contradiction.

For χ of the form $\Box_{\mathcal{C}}^*\psi$, we use the standard argument for PDL (see Kozen and Parikh 1981). This is based on lemmas that parallel Lemmas 6 and 7. The work is somewhat easier than what we do below for sentences of the form $[\alpha]\Box_{\mathcal{C}}^*\psi$, and so we omit these details.

We conclude with the case when χ is a normal form sentence of the form $[\alpha]\Box_{\mathcal{C}}^*\psi \in \Delta$. Assume that $[\alpha]\Box_{\mathcal{C}}^*\psi \in \Delta$. First, suppose that $[\alpha]\Box_{\mathcal{C}}^*\psi \notin U$. Then by Lemma 7, there is a good path from $[U]$ for $\langle\alpha\rangle\Diamond_{\mathcal{C}}^*\neg\psi$. We want to apply Lemma 1 in $\mathcal{F}$ to assert that $\langle\mathcal{F}, [U]\rangle \models \langle\alpha\rangle\Diamond_{\mathcal{C}}^*\neg\psi$. Let k be the length of the good path. For $i \leq k$, $\text{PRE}(\alpha_i) \in U_i$. Now each $\text{PRE}(\alpha_i)$ belongs to Δ by Lemma 5, part (5), and is $< [\alpha]\Box_{\mathcal{C}}^*\psi$. So by induction hypothesis, $\langle\mathcal{F}, [U_i]\rangle \models \text{PRE}(\alpha_i)$. We also need to check that $\langle\mathcal{F}, [U_k]\rangle \models \langle\alpha_k\rangle\neg\psi$. For this, recall from Lemma 5, part (5) that Δ contains

$nf([\alpha_k]\psi) \leq [\alpha_k]\psi$. By Lemma 4, $nf([\alpha_k]\psi) \leq [\alpha_k]\psi < [\alpha]\Box_{\mathcal{C}}^*\psi$. Since the path is good, U_k contains $\langle\alpha_k\rangle\neg\psi$ and hence $\neg[\alpha_k]\psi$. It also must contain the normal form of this, by Lemma 4. So by induction hypothesis, $\langle\mathcal{F}, [U_k]\rangle \models nf(\neg[\alpha_k]\psi)$. By soundness, $\langle\mathcal{F}, [U_k]\rangle \models \langle\alpha_k\rangle\neg\psi$. Now it does follow from Lemma 1 that $\langle\mathcal{F}, [U]\rangle \models \langle\alpha\rangle \Diamond_{\mathcal{C}}^* \neg\psi$.

Going the other way, suppose that $\langle\mathcal{F}, [U]\rangle \models \langle\alpha\rangle \Diamond_{\mathcal{C}}^* \neg\psi$. By Lemma 1, we get a path in $\mathcal{F}$ witnessing this. The argument of the previous paragraph shows that this path is a good path from $[U]$ for $\langle\alpha\rangle \Diamond_{\mathcal{C}}^* \neg\psi$. By Lemma 6, U contains $\langle\alpha\rangle \Diamond_{\mathcal{C}}^* \neg\psi$. This completes the proof.

Theorem 9 (Completeness). *For all φ, $\vdash \varphi$ iff $\models \varphi$. Moreover, this relation is decidable.*

Proof. By Lemma 4, $\vdash \varphi \leftrightarrow nf(\varphi)$. Let φ be consistent. By the Truth Lemma, $nf(\varphi)$ holds at some world in the filtration $\mathcal{F}$. So $nf(\varphi)$ has a model; thus φ has one, too. This establishes completeness. For decidability, note that the size of the filtration is computable in the size of the original φ.

Two Extensions

We briefly mention two extensions of the Completeness Theorem 9. These extensions come from our discussion at the end of section "More on Actions".

First, consider the case of S5 actions. We change our logical system by restricting to these S5 actions, and we add the S5 axioms to our logical system. We interpret this new system on S5 models. It is easy to check that applying an S5 action to an S5 model gives another S5 model. Further, the S5 actions are closed under composition. Finally, if α is an S5 action and $\alpha \rightarrow_A \beta$, then β also is an S5 action. These easily imply the soundness of the new axioms. For completeness, we need only check that if we assume the S5 axioms, then the filtration $\mathcal{F}$ from the previous section has the property that each $\stackrel{A}{\rightarrow}$ is an equivalence relation. This is a standard exercise in modal logic (see, e.g., Fagin et al. (1995), Theorem 3.3.1).

Our second extension concerns the move from actions as we have been working them to actions which change the truth values of atomic sentences. If we make this move, then the axiom of Atomic Permanence is no longer sound. However, it is easy to formulate the relevant axioms. For example, if we have an action α which effects the change $p := p \wedge \neg q$, then we would take an axiom $[\alpha]p \leftrightarrow (\text{PRE}(\alpha) \rightarrow p \wedge \neg q)$. Having made these changes, all of the rest of the work we have done goes through. In this way, we get a completeness theorem for this logic.

Results on Expressive Power

In this section, we present two results which show that adding announcements to modal logic with $\Diamond^*$ adds expressive power as does adding private announcements to modal logic with $\Diamond^*$ and public announcements. To show these results it will

be sufficient to take the set $\mathcal{A}$ of agents to be $\{A, B\}$ and consider only languages contained in a language built-up from the atomic sentences p and q, using $\Diamond_A$, $\Diamond_B$, $\Diamond_A^*$, $\Diamond_B^*$, and $\Diamond_{AB}^*$, and the actions $[\varphi]_A$, $[\varphi]_B$ of announcing φ to A or B privately, and $[\varphi]_{AB}$ the action of announcing φ to A and B publicly. Let $\mathcal{L}_{all}$ stand for this language. We use here the customary notation ($[\varphi]_A$, $[\varphi]_B$, $[\varphi]_{AB}$) for announcements, but $[\varphi]_A$ is simply the action with the Kripke structure $K = \{k\}$ with $\overset{A}{\rightarrow}$ from k to k and $\text{PRE}(k) = \varphi$. We think of $[\varphi]_B$ similarly. $[\varphi]_{AB}$ is the action with the Kripke structure $K = \{k\}$ with $\overset{A}{\rightarrow}$ and $\overset{B}{\rightarrow}$ going from k to k and $\text{PRE}(k) = \varphi$.

We need to define a *rank* $|\varphi|$ on sentences from $\mathcal{L}_{all}$. Let $|p| = 0$ for p atomic, $|\neg\varphi| = |\varphi|$, $|\varphi \wedge \psi| = \max(|\varphi|, |\psi|)$, $|\neg\varphi| = |\varphi|$, $|\Diamond_X \varphi| = 1 + |\varphi|$, for $X = A$ or $X = B$, $|\Diamond_X^* \varphi| = 1 + |\varphi|$ for $X = A$, $X = B$, or $X = AB$, and $|[\varphi]_X\psi| = \max(|\varphi|, |\psi|)$ for $X = A$, $X = B$, or $X = AB$.

First we present a lemma which allows us, in certain circumstances, to do the following: from the existence of a sentence in a language $\mathcal{L}_1$ which is not equivalent to any sentence in a language $\mathcal{L}_0$ infer that there exists a sentence in $\mathcal{L}_1$ not equivalent to any theory in $\mathcal{L}_0$.

Lemma 1. *Let $\mathcal{L}_0$ be a language included in $\mathcal{L}_{all}$, and let ψ be a sentence in $\mathcal{L}_{all}$. Assume that for each n we have models F_n and G_n with some worlds $f_n \in F_n$ and $g_n \in G_n$ such that $\langle F_n, f_n\rangle$ satisfies $\neg\psi$, $\langle G_n, g_n\rangle$ satisfies ψ, and $\langle F_n, f_n\rangle$ and $\langle G_n, g_n\rangle$ agree on all sentences in $\mathcal{L}_0$ of rank $\leq n$. Then $\Diamond_A\psi$ is not equivalent with any theory in $\mathcal{L}_0$.*

Proof. For a sequence of model-world pairs $\langle H_n, h_n\rangle$, $n \in D \subseteq \omega$, we let $\bigoplus_{n\in D}(H_n, h_n)$ be a model-world pair defined as follows. Let h be a new world. Take disjoint copies of the H_n's and add an A-arrow from h to each h_n. All other arrows are within the H_n's and stay the same as in H_n. No atomic sentences are true at h. Atomic sentences true in the worlds belonging to the copy of H_n in $\bigoplus_{n\in D}(H_n, h_n)$ are precisely those true at the corresponding worlds of H_n.

Let F be $\bigoplus_{n\in\omega}(F_n, f_n)$ with the new world denoted by f. Define also F^m, for $m \in \omega$, to be $\bigoplus_{n\in\omega}(H_n, h_n)$ with the new world f^m where $H_m = G_m$, $h_m = g_m$ and for all $n \neq m$, $H_n = F_n$ and $h_n = f_n$

Now assume towards a contradiction that $\Diamond_A\psi$ is equivalent with a theory Φ in $\mathcal{L}_0$. Clearly $\Diamond_A\psi$ fails in $\langle F, f\rangle$. Thus some sentence $\varphi \in \Phi$ fails in $\langle F, f\rangle$. On the other hand, each $\langle F^m, f^m\rangle$ satisfies $\Diamond_A\psi$, whence $\langle F^m, f^m\rangle$ satisfies φ. Let $m_0 = |\varphi|$. The following claim shows that both $\langle F, f\rangle$ and $\langle F^{m_0}, f^{m_0}\rangle$ make φ true or both of them make it false, which leads to a contradiction.

Claim. Let φ be a sentence in Φ of rank $\leq m$. Let H_n, K_n, $n \in D$, with $h_n \in H_n$ and $k_n \in K_n$ be models such that $\langle H_n, h_n\rangle$ and $\langle K_n, k_n\rangle$ agree on sentences in Φ of rank $\leq m$. Then $\langle \bigoplus_n(H_n, h_n), h\rangle$ and $\langle \bigoplus_n(K_n, k_n), k\rangle$ agree on φ.

This claim is proved by induction on complexity of φ. It is clear for atomic sentences. The induction steps for boolean connectives are trivial. A moment of thought gives the induction step for $\Diamond$ and $\Diamond^*$ with various subscripts. It remains to consider the case when $\varphi = [\varphi_1]_A\varphi_2$. (The cases when $\varphi = [\varphi_1]_B\varphi_2$ and $\varphi = [\varphi_1]_{AB}\varphi_2$ are similar.) Fix H_n, K_n, $h_n \in H_n$, $k_n \in K_n$, with $n \in D$, such that $\langle H_n, h_n\rangle$

and $\langle K_n, k_n\rangle$ agree on sentences in Φ of rank $\leq m$. Note that, for each $n \in D$, $\langle H_n, h_n\rangle \models \varphi_1$ if and only if $\langle K_n, k_n\rangle \models \varphi_1$. Let D_1 be the set of all $n \in D$ for which $\langle H_n, h_n\rangle \models \varphi_1$. Let H'_n and K'_n be models obtained by updating H_n and K_n by $[\varphi_1]_A$. By the definition of rank and the fact that $|\varphi_1| \leq m$, we have that $\langle H'_n, h_n\rangle$ and $\langle K'_n, k_n\rangle$ agree on sentences from Φ of rank $\leq m$. Therefore, by our inductive hypothesis

$$\langle \bigoplus_{n\in D_1} H'_n, h\rangle = \varphi_2 \quad \text{iff} \quad \langle \bigoplus_{n\in D_1} K'_n, k\rangle \models \varphi_2.$$

However,

$$\langle \bigoplus_{n} K_n, k\rangle \models \varphi \quad \text{iff} \quad \langle \bigoplus_{n\in D_1} K'_n, k\rangle \models \varphi_2,$$

and

$$\langle \bigoplus_{n} H_n, h\rangle = \varphi \quad \text{iff} \quad \langle \bigoplus_{n\in D_1} H'_n, h\rangle \models \varphi_2$$

and we are done.

Announcements Add Expressive Power to Modal Logic with $\Box^*$

In the result below, there will be only one agent A, and so we omit the letter A from the notation. We let $\mathcal{L}([\,], \Diamond^*)$ be modal logic with announcements (to this A) and $\Diamond^* = \Diamond^*_A$. We also let $\mathcal{L}(\Diamond^*)$ be the obvious sublanguage.

Theorem 2. *There is a sentence of* $\mathcal{L}([\,], \Diamond^*)$ *which cannot be expressed by any set of sentences of* $\mathcal{L}(\Diamond^*)$.

Proof. We show first that $[p]\,\Diamond^+ q = [p]\,\Diamond\,\Diamond^* q$ cannot be expressed by any single sentence of $\mathcal{L}(\Box^*)$. (Incidentally, the same holds for $[p]\,\Diamond^* q$.) Fix a natural number n. We define structures $\mathcal{A} = \mathcal{A}_n$ and $\mathcal{B} = \mathcal{B}_n$ as follows. First $\mathcal{B}$ has $2n + 3$ points arranged cyclically as

$$0 \to 1 \to \cdots \to n \to n + 1 \to -n \to \cdots \to -1 \to 0.$$

For the atomic sentences, we set p true at all points except $n + 1$, and q true only at 0.

The structure $\mathcal{A}$ is a copy of $\mathcal{B}$ with n more points $\overline{1}, \ldots, \overline{n}$ arranged as

$$0 \to \overline{1} \to \cdots \to \overline{n} \to 0.$$

The shape of $\mathcal{A}$ is a figure-8. In both structures, every point is reachable from every point by the transitive closure of the $\to$ relation. At the points $\overline{i}$, p is true and q is false. Notice that $1 \models [p]\,\Diamond^+ q$ in $\mathcal{A}$, but $1 \not\models [p]\,\Diamond^+ q$ in $\mathcal{B}$.

The main technique in the proof is an adaptation of Fraisse-Ehrenfeucht games to the setting of modal logic. Here is a description of the relevant game $G_n(\langle U, u\rangle, \langle V, v\rangle)$. For $n = 0$, *II* immediately wins if the following holds: for all $p \in \mathsf{AtSen}$, $\langle U, u\rangle \models p$ iff $\langle V, v\rangle \models p$. And if u and v differ on some atomic sentence, i immediately wins. Continuing, here is how we define $G_{n+1}(\langle U, u\rangle, \langle V, v\rangle)$. As in the case of G_0, we first check if u and v differ on some atomic sentence. If they do, then i immediately wins. Otherwise, the play continues. Now i can make two types of moves.

1. $\Diamond$-move

i has a choice of playing from U or from V. If i chooses U, then i continues by choosing some u' such that $u \to u'$ in U. Then *II* replies with some $v' \in V$ such that $v \to v'$. Of course, if i had chosen in V, then *II* would have chosen in U. Either way, points u' and v' are determined, and the two players then play $G_n(\langle U, u'\rangle, \langle V, v'\rangle)$.

2. $\Diamond^*$-move

i plays by selecting U (or V, but we ignore this symmetric case below), and then playing some u' (say) reachable from u in the reflexive-transitive closure $\to^*$ of $\to$; *II* responds with a point in the other model, V, which is similarly related to v.

We write $\langle U, u\rangle \sim_n \langle V, v\rangle$ if *II* has a winning strategy in the game $G_n(\langle U, u\rangle, \langle V, v\rangle)$. It is easy to check that by induction on m that if $\langle U, u\rangle \sim_n \langle V, v\rangle$ and $m < n$, then $\langle U, u\rangle \sim_m \langle V, v\rangle$.

Claim 1. If $\langle U, u\rangle \sim_n \langle V, v\rangle$, then for all φ with $|\varphi| \leq n$, $\langle U, u\rangle \models \varphi$ iff $\langle V, v\rangle \models \varphi$.

The proof will be done by induction on φ. Let φ be atomic. Suppose $\langle U, u\rangle \sim_n \langle V, v\rangle$. Then since *II* has a winning strategy, the atomic sentences satisfied by u and v must be the same. So we are done in this case.

The induction steps for the boolean connectives are trivial. For $\Box\varphi$, suppose that $|\Box\varphi| \leq n$, $\langle U, u\rangle \sim_n \langle V, v\rangle$, and $\langle U, u\rangle \models \Box\varphi$. Suppose towards a contradiction that $\langle V, v\rangle \models \Diamond\neg\varphi$. Let v' be such that $v \to v'$ in V and $\langle V, v'\rangle \models \neg\varphi$. Let i begin a play of $G_{n-1}(\langle U, u\rangle, \langle V, v\rangle)$ by choosing to play $v' \in V$. Then *II*'s winning strategy responds with some u' such that $(U, u') \sim_{n-1} (V, v')$. Since $|\varphi| \leq n - 1$, our induction hypothesis implies that $\langle U, u'\rangle \models \Diamond\neg\varphi$. This is a contradiction.

The argument for $\Box^*\varphi$ is similar and we leave it to the reader. The claim is proved.

We return to the models $\mathcal{A}$ and $\mathcal{B}$ described in the beginning of this proof. For $0 \leq i \leq n$, we let $S_i \subseteq \mathcal{A} \times \mathcal{B}$ be the following set

$$\begin{aligned} S_i \quad = \quad & \{(0,0), \ldots, (n,n), (n+1, n+1), (-n,-n), \ldots, (-1,-1)\} \\ & \cup \{(\overline{n}, -1), (\overline{n-1}, -2) \ldots, (\overline{2}, -(n-1)), (\overline{1}, -n)\} \\ & \cup \{(\overline{1}, 1), \ldots, (\overline{n-i}, n-i)\} \end{aligned}$$

In the case of $i = n$, then the last disjunct is empty. Note that $S_0 \supset S_1 \supset \cdots \supset S_n$. Also, for $0 \leq i \leq n$, every point of one structure is related by S_i to some point of the other.

Claim 2. If $0 \leq i \leq n$ and $(a, b) \in S_i$, then $\langle \mathcal{A}, a\rangle \sim_i \langle \mathcal{B}, b\rangle$.

The proof is by induction on i. If $i = 0$, this is due to the fact that pairs in S_0 agree on the atomic formulas. Assume the statement for i, and that $i + 1 \leq n$. Let $(a, b) \in S_{i+1}$. We only need to show that *II* can respond to any play and have the resulting pair belong to S_i. Suppose first that i plays a $\Diamond$-move. Suppose also that $a = b$, so that (a, a) comes from the first subset of S_{i+1}. In this case, we only need to notice that $(a + 1, a + 1) \in S_i$ if $|a| \leq n$, $(-n, -n) \in S_i$ if $a = n + 1$, and $(\overline{1}, 1) \in S_i$ if $a = 0$, since $i < n$. The case of (a, b) from the second subset is similar. Finally, if $(\overline{a}, a)$ belongs to the third subset of S_{i+1}, then $a \leq n - (i + 1) = n - i - 1$. So $a + 1 \leq n - i$, and $(\overline{a + 1}, a + 1)$ belongs to the third subset of S_i. This tells *II* how to play.

We remarked above that each S_i is a total relation. Moreover, each world can be reached from any other one in $\mathcal{A}$ and in $\mathcal{B}$. This implies that if i makes a $\Diamond^*$-move, *II* can respond. This completes the proof of the claim.

It follows that $\langle \mathcal{A}_n, 1\rangle \sim_n \langle \mathcal{B}_n, 1\rangle$. So by Claim 1, for each sentence $\varphi \in \mathcal{L}(\Box^*)$ and all $n \geq |\varphi|$, $\langle \mathcal{A}_n, 1\rangle \models \varphi$ iff $\langle \mathcal{B}_n, 1\rangle \models \varphi$. This shows that $[p] \Diamond^+ q$ cannot be expressed by a single sentence in $\mathcal{L}(\Diamond^*)$. To prove the stronger result as stated in Theorem 2, we only need to quote Lemma 1.

Private Announcements Add Expressive Power

In this section, $\mathcal{L}([\ ]_{AB}, \Diamond^*)$ denotes the set of sentences built from p using $[\varphi]_{AB}$, $\Diamond_A$, $\Diamond_B$, $\Diamond_A^*$, $\Diamond_B^*$, and $\Diamond_{AB}^*$. $\mathcal{L}([\ ]_A, \Diamond_A^*)$ denotes the set built from p using $[\varphi]_A$, $\Diamond_A^*$, and $\Diamond_B$.

Theorem 3. *There is a sentence of* $\mathcal{L}([\ ]_A, \Diamond_A^*)$ *which cannot be expressed by any set of sentences in* $\mathcal{L}([\ ]_{AB}, \Diamond^*)$.

Proof. We consider $\chi \equiv [p]_A \Diamond_A^* \Diamond_B \neg p$.

Let G_n be the following model. We begin with a cycle in $\xrightarrow{A}$:

$$a_1 \xrightarrow{A} a_\infty \xrightarrow{A} b \xrightarrow{A} a_n \xrightarrow{A} a_{n-1} \xrightarrow{A} \cdots \xrightarrow{A} a_2 \xrightarrow{A} a_1 \tag{38.8}$$

We add edges $a_i \xrightarrow{A} b$ for all i (including $i = \infty$), and also $x \xrightarrow{A} a_\infty$ for all x (again including $x = a_\infty$). The only $\xrightarrow{B}$ edge is $a_1 \xrightarrow{B} b$. The atomic sentence p is true at all points except b.

The first thing to note is that after a *private* update of p to A, $\langle G_n, a_i\rangle \models \chi$ for all $i < \infty$. The relevant path is $a_i \xrightarrow{A} \cdots \xrightarrow{A} a_1 \xrightarrow{B} b$; the important point is that since the announcement was private, the edge $a_1 \xrightarrow{B} b$ survives the update. On the other hand, $\langle G_n, a_\infty\rangle \models \neg\chi$. This is because the only way to go from a_∞ to b is to go through b, and the edge $a_\infty \xrightarrow{A} b$ is lost in the update.

Suppose towards a contradiction that χ were equivalent to $\varphi \in \mathcal{L}([\ ]_{AB}, \Diamond^*)$. Let $i = |\varphi|$, and let $n = i + 1$. As we know from our discussion of χ, $\langle G_n, a_n\rangle \models \chi$ and $\langle G_n, a_\infty\rangle \models \neg\chi$. However, this contradicts the claim below.

Claim. Assume that $1 < j \leq n$, $\varphi \in \mathcal{L}([\ ]_{AB}, \Diamond^*)$ and $|\varphi| < j$. Then $\langle G_n, a_j \rangle \models \varphi$ iff $\langle G_n, a_\infty \rangle \models \varphi$.

The proof is by induction on φ. For $\varphi = p$, the result is clear, as are the induction steps for $\neg$ and $\wedge$. For $\Diamond_A \varphi$, suppose that $a_j \models \Diamond_A \varphi$. Either $a_\infty \models \varphi$, in which case $a_\infty \models \Diamond_A \varphi$, or else $a_{j-1} \models \varphi$. In the latter case, by induction hypothesis, $a_\infty \models \varphi$; whence $a_\infty \models \Diamond_A \varphi$. The converse is similar.

The case of $\Diamond_B \varphi$ is trivial: $a_j \models \neg \Diamond_B \varphi$ and $a_\infty \models \neg \Diamond_B \varphi$.

For $\Diamond_A^* \varphi$, note that since we have a cycle (38.8) containing all points, the truth value of $\Diamond_A^* \varphi$ does not depend on the point. The cases of $\Diamond_B^* \varphi$ and $\Diamond_{AB}^* \varphi$ are similar.

For $[\varphi]_{AB} \psi$, assume the result for φ and ψ, and let $|[\varphi]_{AB} \psi| < j$. Then also $|\varphi| < j$ and $|\psi| < j$. Let $H = \{x : x \models \varphi\}$ be the updated model, and recall that $\langle G_n, x \rangle \models [\varphi]_{AB} \psi$ iff $x \in H$ and $\langle H, x \rangle \models \psi$. We have two cases: First, $H = G_n$. Then $\langle G_n, x \rangle \models [\varphi]_{AB} \psi$ iff $\langle G_n, x \rangle \models \psi$. So we are done by the induction hypothesis.

The other case is when there is some $x \notin H$. If $a_k \notin H$ for some $k \geq j$ or for $k = \infty$, then *all* these a_k do not belong to H. In particular, neither a_j nor a_∞ belong. And so both a_j and a_∞ satisfy $[\varphi]_{AB} \psi$. If $b \notin H$, then H is bisimilar to a one-point model. This is because every $a_i \in H$ would have some $\stackrel{A}{\rightarrow}$-successor in H (e.g., a_∞), and there would be no $\stackrel{B}{\rightarrow}$ edges. So we assume $b \in H$. Thus $a_i \notin H$ for some $i < j$. Let k be least so that for $k \leq l \leq \infty$, $a_l \models \varphi$. Then $1 < k \leq j$. Let $A_{\geq k} = \{a_l : k \leq l \leq \infty\}$. The submodels generated by a_j and a_∞ contain the same worlds: all worlds in $A_{\geq k}$ and b. We claim that $(A_{\geq k} \times A_{\geq k}) \cup \{\langle b, b \rangle\}$ is a bisimulation on H. The verification here is easy.

So in H, a_j and a_∞ agree on all sentences in *any* language which is invariant for bisimulation. Now $\mathcal{L}([\ ]_{AB}, \Diamond^*)$ has this property (as do all the languages which we study: they are translatable into infinitary modal logic). In particular, $\langle H, a_j \rangle \models \psi$ iff $\langle H, a_\infty \rangle \models \psi$. This concludes the claim.

We get Theorem 3 directly from the claim, the observation that $\langle G_n, a_n \rangle \models \chi$ and $\langle G_n, a_\infty \rangle \models \neg \chi$, and Lemma 1.

We feel that our two results on expressive power are just a sample of what could be done in this area. We did not investigate the next natural questions: Do announcements with suspicious outsiders extend the expressive power of modal logic with all secure private announcements and common knowledge operators? And then do announcements with common knowledge of suspicion add further expressive power?

Conclusions and Historical Remarks

The work of this paper builds on the long tradition of epistemic logic as well as technical results in other areas. In recent times, one very active arena for work on knowledge is distributed systems, and the main source of work in recent times on knowledge in distributed systems is the book **Reasoning About Knowledge** (Fagin et al. 1995) by Fagin, Halpern, Moses, and Vardi. We depart from Fagin et al. (1995) by introducing the new operators for epistemic actions, and by doing without

temporal logic operators. In effect, our Kripke models are simpler, since they do not incorporate all of the runs of a system; the new operators can be viewed as a compensation for that. We have not made a detailed comparison of our work with the large body of work on knowledge on distributed systems, and such a comparison would require both technical and conceptual results. On the technical side, we suspect that neither framework is translatable into the other. One way to show this would be by expressivity results. Perhaps another way would use complexity results. In this direction, we note that Halpern and Vardi (1989) examines 96 logics of knowledge and time. Thirty-two of these contain common knowledge operators, and of these, all but twelve of these are undecidable. But overall, our logics are based on differing conceptual points and intended applications, and so we are confident that they differ.

As far as we know, the first paper to study the interaction of communication and knowledge in a formal setting is Plaza's paper "Logics of Public Communications" (Plaza 1989). As the title suggests, the epistemic actions studied are announcements to the whole group, as in our α and α'. Perhaps the main result of the paper is a completeness theorem for the logic of public announcements and knowledge. This result is closely related to a special case of our Theorem 5. The difference is that Plaza restricts attention to the case when all of the accessibility relations are equivalence relations. Incidentally, Plaza's proof involves a translation to multi-modal logic, just as ours does. In addition to this, Plaza (1989) contains a number of results special to the logic of announcements which we have not generalized, and it also studies an extension of the logic with non-rigid constants.

Other predecessors to this paper are the papers of Gerbrandy (1999a,b) and Gerbrandy and Groeneveld (1997). These study epistemic actions similar to our β, where an announcement is made to set of agents in a private way with no suspicions. They presented a logical system which included the common knowledge operators. An important result is that all of the reasoning in the original Muddy Children scenario can be carried out in their system. This shows that in order to get a formal treatment of the problem, one need not posit models which maintain histories. They did not obtain the completeness/decidability result for their system, but it would be the version of Theorem 9 restricted to actions which are compositions of private announcements. So it follows from our work that all of the reasoning in the Muddy Children can be carried out in a decidable system.

We should mention that the systems studied in Gerbrandy (1999a,b) and Gerbrandy and Groeneveld (1997) differ from ours in that they are variants of dynamic logic rather than propositional logic. That is, announcements are particular types of *programs* as opposed to modalities. This is a natural move, and although we have not followed it in this paper, we have carried out a study of expressive power issues of various fragments of a dynamic logic with announcement operators. We have shown, for example, that the dynamic logic formulations are more expressive than the purely propositional ones. Details on this will appear in a forthcoming paper.

Incidentally, the semantics in Gerbrandy (1999a,b) and Gerbrandy and Groeneveld (1997) use non-wellfounded sets. In other words, they work with models modulo bisimulation. The advantages of moving from these to arbitrary Kripke

models are that the logic can be used by those who do not know about non-wellfounded sets, and also that completeness results are slightly stronger with a more general semantics. The relevant equivalence of the two semantics is the subject of the short note (Moss 1999).

The following are the new contributions of this paper:

1. We formulated a logical system with modalities corresponding to intuitive group-level epistemic actions. These actions include natural formalizations of announcements such as γ and δ, which allow various types of suspicion by outsiders. Our apparatus also permits us to study epistemic actions which apparently have not yet been considered in this line of work, such as actions in which nothing actually happens but one agent suspects that a secret communication took place.
2. We formulated a logical system with these modalities and with common knowledge operators for all groups. Building on the completeness of PDL and using a bit of term rewriting theory, we axiomatized the validities in our system.
3. We obtained some results on expressive power: in the presence of common knowledge operators, it is not possible to translate away public announcements, and in our framework, private announcements add expressive power to public ones.

Appendix: The Lexicographic Path Order

In this appendix, we give the details on the *lexicographic path ordering* (*LPO*), both in general and in connection with $\mathcal{L}([\alpha])$ and $\mathcal{L}([\alpha], \Box^*)$.

Fix some many-sorted signature Σ of terms. In order to define the LPO $<$ on the Σ-terms, we must first specify a well-order $<$ on the set of function symbols of Σ. The LPO determined by such choices is the smallest relation $<$ such that:

(LPO1) If $(t_1, \ldots, t_n) < (s_1, \ldots, s_n)$ in the lexicographic ordering on n-tuples, and if $t_j < f(s_1, \ldots, s_n)$ for $1 \leq j \leq n$, then $f(t_1, \ldots, t_n) < f(s_1, \ldots, s_n)$.

(LPO2) If $t \leq s_i$ for some i, then $t < f(s_1, \ldots, s_n)$.

(LPO3) If $g < f$ and $t_i < f(s_1, \ldots, s_n)$ for all $i \leq m$, then $g(t_1, \ldots, t_m) < f(s_1, \ldots, s_n)$.

Here is how this is applied in this paper. We shall take two sorts: *sentences* and *actions*. Our signature contains the usual sentence-forming operators p (for $p \in \mathsf{AtSen}$) $\neg$, $\wedge$, and $\Box_A$ for all $A \in \mathcal{A}$. Here each p is 0-ary, $\neg$ and $\Box_A$ are unary, and $\wedge$ is binary. We also have an operator *app* taking actions and sentences to sentences. We think of $app(\psi, \alpha)$ as merely a variation on $[\alpha]\psi$. (The order of arguments to *app* is significant.) We further have a binary operator $\circ$ on actions. (This is a departure from the treatment of this paper, since we used $\circ$ as a metalinguistic abbreviation instead of as a formal symbol. It will be convenient to make this change because this leads to a smoother treatment of the Composition Axiom.) Finally, for each finite Kripke frame K over $\mathcal{L}([\alpha])$ and each $1 \leq i \leq |K|$, we have a symbol F^i_K taking $|K|$ sentences and returning an action.

Each sentence φ has a formal version $\overline{\varphi}$ in this signature, and each action α also has a formal version $\overline{\alpha}$. These are defined by the recursion which is obvious except for the clauses

$$\begin{array}{ll} \overline{[\alpha]\varphi} & = app(\overline{\varphi}, \overline{\alpha}) \\ \overline{\alpha} & = \mathsf{F}^i_K(\overline{\mathrm{PRE}(k_1)}, \ldots, \overline{\mathrm{PRE}(k_n)}) \end{array}$$

Here $\alpha = \langle K, k_i, \mathrm{PRE}\rangle$ with $K = \{k_1, \ldots, k_n\}$ in some specified order. However, outside of the proof of Proposition 2 we shall not explicitly mention the formal versions at all, since they are harder to read than the standard notation.

We must also first fix a wellfounded relation $<$ on the function symbols. We set *app* to be greater than all other function symbols. In all other cases, distinct function symbols are unordered.

Theorem 1 (Kamin and Levy 1980; Dershowitz 1982). *Let $<$ be an LPO on Σ-terms.*

1. *$<$ is transitive.*
2. *$<$ has the* subterm property*: if t is a proper subterm of u, then $t < u$.*
3. *$<$ is* monotonic *(it has the* replacement property*): if $y < x_i$ for some i, then*

$$f(x_1, \ldots, y, \ldots x_n) < f(x_1, \ldots, x_i, \ldots x_n).$$

4. *$<$ is wellfounded.*
5. *Consider a term rewriting system every rule of which of the form $l \rightsquigarrow r$ with $r < l$. Then the system is* terminating*: there are no infinite sequences of rewritings.*

Proof. Here is a sketch for part (1): We check by induction on the construction of the least relation $<$ that if $s < t$, then for all u such that $t < u$, $s < u$. For this, we use induction on the term u. We omit the details. Further, (2) follows easily from (1) and (LPO2), and (3) from (LPO1), (1) and (2). Moreover, (5) follows easily from (4) and (3), since the latter implies that any replacement according to the rewrite system results in a smaller term in the order $<$.

Here is a proof of of the wellfoundedness property (4), taken from on Buchholz (1995). (We generalized it slightly from the one-sorted to the many-sorted setting and from the assumption that $<$ is a finite linear order on Σ to the assumption that $<$ is any wellfounded order.)

Let W be the set of terms t such that the order $<$ is wellfounded below t. W is then itself wellfounded under $<$. So for all n, W^n is wellfounded under the induced lexicographic order. We prove by induction on the given wellfounded relation on function symbols of Σ that for all n-ary f, $f[W^n] \subseteq W$. So assume that for $g < f$, say with arity m, $g[W^m] \subseteq W$. We check this for f by using induction on W^n. Fix $\vec{s} \in W^n$, and assume that whenever $\vec{u} < \vec{s}$ in W^n, that $f(\vec{u}) \in W$. We prove that $f(\vec{s}) \in W$ by checking that for all t such that $t < f(\vec{s})$, $t \in W$. And *this* is done by induction on the structure of t. If $t = f(\vec{u}) < f(\vec{s})$ via (LPO1), then $\vec{u} < \vec{s}$ lexicographically, and each

$u_i < f(\vec{s})$. This last point implies that $\vec{u} \in W^n$ by induction hypothesis on t, so $t \in W$ by induction hypothesis on W^n. If $t \leq s_i$ so that $t < f(\vec{s})$ via (LPO2), then $t \in W$ by definition of W. And if $t = g(u_1, \ldots, u_m) < f(\vec{s})$ via (LPO3), then $g < f$ and each $u_i < f(\vec{s})$. By induction hypothesis on t, each $u_i \in W$. So by induction hypothesis on f, $g(\vec{u}) \in W$.

Now that we know that each f takes tuples in W^n to elements of W, it follows by induction on terms that all terms belong to W.

For more on the LPO, its generalizations and extensions, see the surveys Dershowitz (1982) and Plaisted (1985).

Proposition 2. *Consider the LPO $<$ on $\mathcal{L}([\alpha], \Box^*)$ defined above.*

1. *If $\alpha \xrightarrow{*} \beta$, then $\text{PRE}(\beta) < \alpha$.*
2. *If $\alpha \xrightarrow{*} \beta$, then $[\beta]\psi < [\alpha]\Box^*_{\mathcal{C}}\psi$.*
3. *$\text{PRE}(\alpha) \to p < [\alpha]p$.*
4. *$\text{PRE}(\alpha) \to \neg[\alpha]\psi < [\alpha]\neg\psi$.*
5. *$[\alpha]\psi \wedge [\alpha]\chi < [\alpha](\psi \wedge \chi)$.*
6. *$\text{PRE}(\alpha) \to \bigwedge\{\Box_A[\beta]\psi : \alpha \xrightarrow{A} \beta\} < [\alpha]\Box_A\psi$.*
7. *$[\alpha \circ \beta]\varphi < [\alpha][\beta]\psi$.*

In particular, for all rules $\varphi \rightsquigarrow \psi$ of the rewriting system $\mathcal{R}^*$, $\psi < \varphi$.

Proof. Part (1) holds because we regard α as a term $\alpha = \mathsf{F}^i_K(\overline{\gamma_1}, \ldots, \overline{\gamma_n})$, for some frame K and i. So whenever $\alpha \xrightarrow{*} \beta$, each $\text{PRE}(\beta)$ is a proper subterm of α.

Here is the argument for part (2): We need to see that $app(\psi, \overline{\beta}) < app(\Box^*_{\mathcal{C}}\psi, \overline{\alpha})$. Now lexicographically, $(\psi, \overline{\beta}) < (\Box^*_{\mathcal{C}}\psi, \overline{\alpha})$. So we only need to know that $\overline{\beta} < app(\Box^*_{\mathcal{C}}\psi, \overline{\alpha})$. Let $\alpha = \mathsf{F}^i_K(\overline{\gamma_1}, \ldots, \overline{\gamma_n})$. Now according to Eq. (38.2) in section "Syntax", $\overline{\beta}$ is $\mathsf{F}^j_K(\overline{\gamma_1}, \ldots, \overline{\gamma_n})$, for the same K and $\gamma_1, \ldots, \gamma_n$ but perhaps for $j \neq i$. Then it is clear by (LPO2) that $\overline{\gamma_i} < app(\Box^*_{\mathcal{C}}\psi, \overline{\alpha})$ for all i. So by (LPO3), $\overline{\beta} < app(\Box^*_{\mathcal{C}}\psi, \overline{\alpha})$.

The remaining parts are similar.

A *normal form* in a rewriting system is a sentence which cannot be rewritten in the system. Of course, we are interested in the systems $\mathcal{R}$ and $\mathcal{R}^*$ from sections "A Logic for $\mathcal{L}([\alpha])$" and "A Logic for $\mathcal{L}([\alpha], \Box^*)$", respectively. It follows from the wellfoundedness of $<$ that for every φ there is a normal form $nf(\varphi) \leq \varphi$ obtained by rewriting φ in some arbitrary fashion until a normal form is reached.

Lemma 3. *A sentence $\varphi \in \mathcal{L}([\alpha])$ is a normal form of $\mathcal{R}^*$ iff φ is a modal sentence (i.e., iff φ contains no actions). Moreover, the rule $[\alpha][\beta]\varphi \rightsquigarrow [\alpha \circ \beta]\varphi$ is not needed to reduce φ to normal form. So for $\mathcal{L}([\alpha])$, $\mathcal{R}$ has the same normal forms as $\mathcal{R}^*$.*

A sentence $\varphi \in \mathcal{L}([\alpha], \Box^*)$ is a normal form of $\mathcal{R}^*$ iff φ is built from atomic sentences using $\neg$, $\wedge$, $\Box_A$, and $\Box^*_{\mathcal{B}}$, or if φ is of the form $[\alpha]\Box^*_{\mathcal{B}}\psi$, where α is a normal form action, and ψ is a normal form. An action α is a normal form if whenever $\alpha \xrightarrow{*} \beta$, then $\text{PRE}(\beta)$ is a normal form.

Proof. It is immediate that every modal sentence is a normal form in $\mathcal{L}([\alpha])$, that every $[\alpha]\Box^*_C\varphi$ is a normal form in $\mathcal{L}([\alpha], \Box^*)$, and that if each $\text{PRE}(\beta)$, with $\alpha \xrightarrow{*} \beta$, is a normal form, then α is a normal form action. Going the other way, we check that if $\varphi \in \mathcal{L}([\alpha])$, $[\alpha]\varphi$ is *not* a normal form. So we see by an easy induction that the normal forms of $\mathcal{L}([\alpha])$ are exactly the modal sentences. We also argue by induction for $\mathcal{L}([\alpha], \Box^*)$, and we note that every $[\alpha][\beta]\varphi$ is *not* a normal form, using the rule $[\alpha][\beta]\varphi \rightsquigarrow [\alpha \circ \beta]\varphi$.

One fine point concerning $\mathcal{R}$ and our work in section "A Logic for $\mathcal{L}([\alpha])$" is that to reduce sentences of $\mathcal{L}([\alpha])$ to normal form we may restrict ourselves to rewriting sentences which are not subterms of actions. This simplification accounts for the differences between parallel results of sections "A Logic for $\mathcal{L}([\alpha])$" and "A Logic for $\mathcal{L}([\alpha], \Box^*)$".

Acknowledgements We thank Jelle Gerbrandy and Rohit Parikh for useful conversations on this work. An earlier version of this paper was presented at the 1998 Conference on Theoretical Aspects of Rationality and Knowledge.

References

Baltag, A. (1999). *A logic of epistemic actions, ms*. Amsterdam: CWI.

Buchholz, W. (1995). Proof-theoretic analysis of termination proofs. *Annals of Pure and Applied Logic, 75*, 57–65.

Dershowitz, N. (1982). Orderings for term-rewriting systems. *Theoretical Computer Science, 17*, 279–301.

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning about knowledge*. Cambridge: MIT.

Gerbrandy, J. (1999a). Dynamic epistemic logic. In L. S. Moss, et al. (Eds.), *Logic, language, and information* (Vol. 2). Stanford: CSLI Publications, Stanford University.

Gerbrandy, J. (1999b). *Bisimulations on planet Kripke*. Ph.D. dissertation, University of Amsterdam.

Gerbrandy, J., & Groeneveld, W. (1997). Reasoning about information change. *Journal of Logic, Language, and Information, 6*, 147–169.

Halpern, J. Y., & Vardi, M. Y. (1989). The complexity of reasoning about knowledge and time. I. Lower bounds. *Journal of Computer and System Sciences, 38*(1), 195–237.

Kamin, S., & Levy, J. J. (1980). *Two generalizations of the recursive path orderings*. Unpublished note, Department of Computer Science, University of Illinois, Urbana.

Kozen, D., & Parikh, R. (1981). An elementary proof of the completeness of PDL. *Theoretical Computer Science, 14*, 113–118.

Moss, L. S. (1999). From hypersets to Kripke models in logics of announcements. In J. Gerbrandy, et al. (Eds.), *JFAK. Essays dedicated to Johan van Benthem on the occasion of his 50th birthday*. Amsterdam: Vossiuspers, Amsterdam University Press.

Plaisted, D. A. (1985). Termination orderings. In D. Gabbay, et al. (Eds.), *Handbook of logic in artificial intelligence and logic programming* (Vol. I, pp. 273–364).

Plaza, J. (1989). Logics of public communications. In *Proceedings of the 4th international symposium on methodologies for intelligent systems*, Charlotte.

Chapter 39
A Qualitative Theory of Dynamic Interactive Belief Revision

Alexandru Baltag and Sonja Smets

Introduction

This paper contributes to the recent and on-going work in the logical community (Aucher 2003; Baltag and Sadrzadeh 2006; Baltag and Smets 2006a,b,c; van Benthem 2007; van Ditmarsch 2005) on dealing with mechanisms for belief revision and update within the Dynamic-Epistemic Logic (DEL) paradigm. DEL originates in the work of Gerbrandy and Groeneveld (1997) and Gerbrandy (1999), anticipated by Plaza in (1989), and further developed by numerous authors Baltag et al. (1998), Gerbrandy (1999), van Ditmarsch (2000, 2002), Baltag (2002), Kooi (2003), Baltag and Moss (2004), and van Benthem et al. (2006a,b) etc. In its standard incarnation, as presented e.g., in the recent textbook by van Ditmarsch et al. (2007), the DEL approach is particularly well fit to deal with *complex multi-agent learning actions* by which groups of interactive agents update their beliefs (including *higher-level beliefs* about the others' beliefs), *as long as the newly received information is consistent with the agents' prior beliefs*. On the other hand, the classical AGM theory and its more recent extensions have been very successful in dealing with the problem of *revising one-agent, first-level (factual) beliefs when they are contradicted by new information*. So it is natural to look for a way to combine these approaches.

A. Baltag (✉) • S. Smets
ILLC, University of Amsterdam, The Netherlands
e-mail: thealexandrubaltag@gmail.com; S.J.L.Smets@uva.nl

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_39

We develop here a notion of *doxastic actions*,[1] general enough to cover most examples of multi-agent communication actions encountered in the literature, but also flexible enough to deal with *(both static and dynamic) belief revision*, and in particular to *implement various "belief-revision policies" in a unified setting*. Our approach can be seen as a natural extension of the work in Baltag and Moss (2004) and Baltag et al. (1998) on "epistemic actions", incorporating ideas from the AGM theory along the lines pioneered in Aucher (2003) and van Ditmarsch (2005), but using a *qualitative* approach based on *conditional beliefs*, in the line of Stalnaker (1968), Bonanno (2005), Board (2002), and van Benthem (2007).

Our paper assumes the general distinction, made in Baltag and Smets (2006a); van Benthem (2007); van Ditmarsch (2005), between *"dynamic" and "static" belief revision*. It is usually acknowledged that the classical AGM theory in Alchourrón et al. (1985) and Gärdenfors (and embodied in our setting by the *conditional belief* operators $B_a^P Q$) is indeed "static", in the sense that it captures *the agent's changing beliefs about an unchanging world*. But in fact, when we take into account all the higher-level beliefs, the "world" (that these higher-level beliefs are about) includes all agent's (real) beliefs.[2] Thus, such a world is *always changed by our changes of beliefs*! So we can better understand a belief conditional on P as capturing the agent's beliefs *after revising with P* about the state of the world *before the revision*: the statement $B_a^P Q$ says that, *if agent a would learn P, then she would come to believe that Q was the case (before the learning)*. In contrast, "dynamic" belief revision uses dynamic modalities to capture the agent's revised beliefs about the world *as it is after revision*: $[!P]B_a Q$ says that *after learning P, agent a would come to believe that Q is the case (in the world after the learning)*. The standard alternative (Katsuno and Mendelzon 1992) to the AGM theory calls this *belief update*, but like the AGM approach, it only deals with "first-level" (factual) beliefs from a non-modal perspective, neglecting any higher-order "beliefs about beliefs". As a result, *it completely misses the changes induced* (in our own or the other agents' epistemic-doxastic states) *by the learning actions themselves* (e.g., the learning of a Moore sentence, see the third section on ""Dynamic" Belief Revision"). This is reflected in the acceptance in Katsuno and Mendelzon (1992) of the AGM "Success Axiom": in dynamic notation, this is the axiom $[!P]B_a P$ (which cannot accommodate Moore sentences). Instead, Katsuno and Mendelzon (1992) exclusively concentrate on the possible changes of (ontic) facts that may have occurred during our learning (but *not due to our learning*). In contrast, our approach to belief update (following the DEL tradition) may be thought of as "dual" to the one in Katsuno and Mendelzon

[1] Or "doxastic events", in the terminology of van Benthem (2007).

[2] To verify that a higher-level belief about another belief is "true" we need to check the content of that higher-level belief (i.e., the existence of the second, lower-level belief) against the "real world". So the real world has to include the agent's beliefs.

(1992): we completely neglect here the ontic changes,[3] considering only the changes induced by *"purely doxastic" actions* (learning by observation, communication, etc.).

Our formalism for "static" revision can best be understood as a modal-logic implementation of the well-known view of belief revision in terms of *conditional reasoning* (Stalnaker 1968, 2006). In Baltag and Smets (2006a,c), we introduced two equivalent semantic settings for conditional beliefs in a multi-agent epistemic context (*conditional doxastic models* and *epistemic plausibility models*), taking the first setting as the basic one. Here, we adopt the second setting, which is closer to the standard semantic structures used in the literature on modeling belief revision (Board 2002; Friedmann and Halpern 1994; Grove 1988; Spohn 1988; Stalnaker 2006; van Benthem 2007, 2004). We use this setting to define notions of *knowledge* K_aP, *belief* B_aP and *conditional belief* B_a^QP. Our concept of "knowledge" is the standard S5-notion, partition-based and fully introspective, that is commonly used in Computer Science and Economics, and is sometimes known as "Aumann knowledge", as a reference to Aumann (1999). The conditional belief operator is a way to "internalize", in a sense, the "static" (AGM) belief revision within a modal framework: saying that, at state s, agent a believes P conditional on Q is a way of saying that Q belongs to a's revised "theory" (capturing her revised beliefs) after revision with P (of a's current theory/beliefs) at state s. Our conditional formulation of "static" belief revision is close to the one in Stalnaker (1968), Ryan and Schobbens (1997), Board (2002), Bonanno (2005), and Rott (1989). As in Board (2002), the preference relation is assumed to be well-preordered; as a result, the logic CDL of conditional beliefs is equivalent to the strongest system in Board (2002).

We also consider other modalities, capturing other "doxastic attitudes" than just knowledge and conditional belief. The most important such notion expresses a form of "weak (non-introspective) knowledge" $\Box_aP$, first introduced by Stalnaker in his modal formalization (Stalnaker 1968, 2006) of Lehrer's *defeasibility analysis of knowledge* (Lehrer 1990; Lehrer and Paxson 1969). We call this notion *safe belief*, to distinguish it from our (Aumann-type) concept of knowledge. Safe belief can be understood as belief that is *persistent under revision with any true information*. We use this notion to give a new solution to the so-called "Paradox of the Perfect Believer". We also solve the open problem posed in Board (2002), by providing a *complete axiomatization of the "static" logic* $K\Box$ *of conditional belief, knowledge and safe belief*. In a forthcoming paper, we apply the concept of safe belief to Game Theory, improving on Aumann's epistemic analysis of backwards induction in games of perfect information.

Moving thus on to *dynamic belief revision*, the first thing to note is that (unlike the case of "static" revision), *the doxastic features of the actual "triggering event"* that induced the belief change *are essential* for understanding this change (as a "dynamic

[3] But our approach can be easily modified to incorporate ontic changes, along the lines of van Benthem et al. (2006b).

revision", i.e., in terms of the revised beliefs about the state of the world after revision). For instance, our beliefs about *the current situation after* hearing a *public* announcement (say, of some *factual* information, denoted by an atomic sentence p) are different from our beliefs after receiving a *fully private* announcement with the same content p. Indeed, in the public case, we come to believe that p is now *common knowledge* (or at least *common belief*). While, in the private case, we come to believe that the content of the announcement forms now our *secret knowledge*. So the agent's *beliefs about the learning actions* in which she is currently engaged affect the way she updates her previous beliefs.

This distinction is irrelevant for "static" revision, since e.g., in both cases above (public as well as private announcement) we learn the same thing about the situation that existed *before the learning*: our beliefs about that past situation will change in the same way in both cases. More generally, our beliefs about the "triggering action" are irrelevant, as far as our "static" revision is concerned. This explains a fact observed in van Benthem (2007), namely that by and large, the standard literature on belief revision (or belief update) *does not usually make explicit the doxastic events that "trigger" the belief change* (dealing instead only with types of abstract operations on beliefs, such as update, revision and contraction etc). The reason for this lies in the "static" character of AGM revision, as well as its restriction (shared with the "updates" of Katsuno and Mendelzon 1992) to one-agent, first-level, factual beliefs.

A "truly dynamic" logic of belief revision has to be able to capture the *doxastic-epistemic features (e.g., publicity, complete privacy etc.) of specific "learning events"*. We need to be able to model the agents' "dynamic beliefs", i.e., their *beliefs about the learning action itself*: the *appearance* of this action (while it is happening) to each of the agents. In Baltag and Moss (2004), it was argued that a natural way to do this is to use *the same type of formalism that was used to model "static" beliefs*: *epistemic actions should be modeled in essentially the same way as epistemic states*; and this common setting was taken there to be given by *epistemic Kripke models*.

A similar move is made here in the context of our richer doxastic-plausibility structures, by introducing *plausibility pre-orders on actions* and developing a notion of "action plausibility models", that extends the "epistemic action models" from Baltag and Moss (2004), along similar lines to (but without the quantitative features of) the work in Aucher (2003) and van Ditmarsch (2005).

Extending to (pre)ordered models the corresponding notion from Baltag and Moss (2004), we introduce an operation of *product update* of such models, based on the *anti-lexicographic order* on the product of the state model with the action model. The simplest and most natural way to define a connected pre-order on a Cartesian product from connected pre-orders on each of the components is to use either the *lexicographic* or the *anti-lexicographic* order. Our choice is the second, which we regard as the *natural generalization of the AGM theory*, giving *priority to incoming information* (i.e., to "actions" in our sense). This can also be thought of as a generalization of the so-called *"maximal-Spohn" revision*. We call this type of update rule the *"Action-Priority" Update*. The intuition is that the beliefs encoded in the action model express *the "incoming" changes of belief*, while the state model

only captures that *past beliefs*. One could say that the new "beliefs about actions" are *acting* on the prior "beliefs about states", producing the updated (posterior) beliefs. This is embedded in the Motto that we give in the paragraph on "Action Models" in the third section, the Motto is: "*beliefs about changes encode (and induce) changes of beliefs*".

By abstracting away from the quantitative details of the plausibility maps when considering the associated *dynamic logic*, our approach to dynamic belief revision is in the spirit of the one in van Benthem (2007): instead of using "graded belief" operators as in e.g., Aucher (2003) and van Ditmarsch (2005), or probabilistic modal logic as in Kooi (2003), both our account and the one in van Benthem (2007) concentrate on the simple, qualitative language of *conditional beliefs, knowledge and action modalities* (to which we add here the *safe belief* operator). As a consequence, we obtain *simple, elegant, general logical laws of dynamic belief revision*, as natural generalizations of the ones in van Benthem (2007). These "reduction laws" give a *complete axiomatization of the logic of doxastic actions*, "reducing" it to the "static" logic $K\Box$. Compared both to our older axiomatization in Baltag and Smets (2006c) and to the system in Aucher (2003), one can easily see that the introduction of the safe belief operator leads to a major simplification of the reduction laws.

Our qualitative logical setting (in this paper and in Baltag and Smets 2006a,b,c), as well as the closely related setting in van Benthem (2007), are conceptually very different from the more "quantitative" approaches to dynamic belief revision taken in (Aucher 2003; van Ditmarsch 2005; van Ditmarsch and Labuschagne 2007), approaches based on "degrees of belief" given by ordinal plausibility functions. This is not just a matter of interpretation, but it makes a difference for the choice of dynamic revision operators. Indeed, the update mechanisms proposed in Spohn (1988), Aucher (2003), and van Ditmarsch (2005) are essentially quantitative, using various binary functions in transfinite ordinal arithmetic, in order to compute the degree of belief of the output-states in terms of the degrees of the input-states and the degrees of the actions. This leads to an increase in complexity, both in the computation of updates and in the corresponding logical systems. Moreover, there seems to be no canonical choice for the arithmetical formula for updates, various authors proposing various formulas. No clear intuitive justification is provided to any of these formulas, and we see no transparent reason to prefer one to the others. In contrast, classical (AGM) belief revision theory is a qualitative theory, based on natural, intuitive postulates, of great generality and simplicity.

Our approach retains this qualitative flavor of the AGM theory, and aims to build a theory of "dynamic" belief revision of equal simplicity and naturality as the classical "static" account. Moreover (unlike the AGM theory), it aims to provide a *"canonical"* choice for a dynamic revision operator, given by our "Action Priority" update. This notion is a *purely qualitative one*,[4] based on a *simple, natural relational*

[4]One could argue that our plausibility pre-order relation is equivalent to a quantitative notion (of ordinal degrees of plausibility, such as in Spohn (1988)), but unlike in Aucher (2003) and van

definition. From a *formal point of view*, one might see our choice of the anti-lexicographic order as *just one of the many possible options* for developing a belief-revision-friendly notion of update. As already mentioned, it is a generalization of the "maximal-Spohn" revision, already explored in van Ditmarsch (2005) and Aucher (2003), among many other possible formulas for combining the "degrees of belief" of actions and states. But here we justify our option, arguing that our *qualitative interpretation of the plausibility order makes this the only reasonable choice*.

It may seem that by making this choice, we have confined ourselves to *only one of the bewildering multitude of "belief revision policies"* proposed in the literature by Spohn (1988), Rott (1989), Segerberg (1998), Aucher (2003), van Ditmarsch (2005), van Benthem (2004), and van Benthem (2007). But, as argued below, *this apparent limitation is not so limiting after all*, but can instead be regarded as an *advantage*: the power of the "action model" approach is reflected in the fact that *many different belief revision policies* can be recovered as *instances of the same type of update operation*. In this sense, our approach can be seen as a *change of perspective*: the diversity of possible revision policies is replaced by the diversity of possible action models; the differences are now viewed as *differences in input, rather than having different "programs"*. For a computer scientist, this resembles "Currying" in lambda-calculus: if every "operation" is encoded as an input-term, then *one operation* (functional application) *can simulate all operations*.[5] In a sense, this is nothing but the idea of Turing's universal machine, which underlies universal computation.

The title of our paper is a paraphrase of Oliver Board's "Dynamic Interactive Epistemology" (Board 2002), itself a paraphrase of the title ("Interactive Epistemology") of a famous paper by Aumann (1999). We interpret the word "interactive" as referring to the *multiplicity of agents* and the *possibility of communication*. Observe that "interactive" does not necessarily imply "dynamic": indeed, Board and Stalnaker consider Aumann's notion to be "static" (since it doesn't accommodate any non-trivial belief revision). But even Board's logic, as well as Stalnaker's (2006), are "static" in our sense: they cannot directly capture the effect of learning *actions* (but can only express "static" conditional beliefs). In contrast, our DEL-based approach has all the "dynamic" features and advantages of DEL: in addition to "simulating" a range of individual belief-revision policies, it can deal with an even wider range of *complex types of multi-agent learning and communication actions*. We thus think it is realistic to expect that, *within its own natural limits*,[6] our Action-Priority Update Rule could play the role of a *"universal machine" for qualitative dynamic interactive belief-revision*.

Ditmarsch (2005) the way belief update is defined in our account does not make any use of the ordinal "arithmetic" of these degrees.

[5]Note that, as in untyped lambda-calculus, the input-term encoding the operation (i.e., our "action model") and the "static" input-term to be operated upon (i.e., the "state model") are essentially *of the same type*: epistemic plausibility models for the same language (and for the same set of agents).

[6]E.g., our update cannot deal with "forgetful" agents, since "perfect recall" is in-built. But finding out what exactly are the "natural limits" of our approach is for now an open problem.

"Static" Belief Revision

Using the terminology in van Benthem (2007) and Baltag and Smets (2006a,b,c, 2007a), "static" belief revision is about *pre-encoding potential belief revisions as conditional beliefs*. A conditional belief statement $B_a^P Q$ can be thought of as expressing a "doxastic predisposition" or a "plan of doxastic action": the agent is determined to believe that Q was the case, if he learnt that P was the case. The semantics for conditional beliefs is usually given in terms of plausibility models (or equivalent notions, e.g., "spheres", "onions", ordinal functions etc.) As we shall see, both *(Aumann, S5-like) knowledge* and *simple (unconditional) belief* can be defined in terms of conditional belief, which itself could be defined in terms of a *unary belief-revision operator*: $*_a P$ captures *all the revised beliefs* of agent a after revising (her current beliefs) with P.

In addition, we introduce a *safe belief* operator $\Box_a P$, meant to express a weak notion of "defeasible knowledge" (obeying the laws of the modal logic $S4.3$). This concept was defined in Stalnaker (2006) and Board (2002) using a higher-order semantics (quantifying over conditional beliefs). But this is in fact equivalent to a first-order definition, as the Kripke modality for the (converse) plausibility relation. This observation greatly simplifies the task of completely axiomatizing the logic of safe belief and conditional beliefs: indeed, our proof system $K\Box$ below is a solution to the open problem posed in Board (2002).

Plausibility Models: The Single Agent Case

To warm up, we consider first the case of only *one agent*, a case which fits well with the standard models for belief revision.

A *single-agent plausibility frame* is a structure $(S, \leq)$, consisting of a set S of "states" and a "well-preorder" $\leq$, i.e., a reflexive, transitive binary relation on S such that *every non-empty subset has minimal elements*. Using the notation

$$\mathrm{Min}_{\leq} P := \{s \in P : s \leq s' \text{ for all } s' \in P\}$$

for the set of $\leq$-minimal elements of P, the last condition says that: For every set $P \subseteq S$, if $P \neq \emptyset$ then $\mathrm{Min}_{\leq} P \neq \emptyset$.

The usual reading of $s \leq t$ is that "state s is *at least as plausible* as state t". We keep this reading for now, though we will later get back to it and clarify its meaning. The "minimal states" in $\mathrm{Min}_{\leq} P$ are thus the "most plausible states" satisfying proposition P. As usual, we write $s < t$ iff $s \leq t$ but $t \not\leq s$, for the *"strict" plausibility relation* (s is *more plausible* than t). Similarly, we write $s \cong t$ iff both $s \leq t$ and $t \leq s$, for the *"equi-plausibility"* (or *indifference*) relation (s and t are *equally plausible*).

S-propositions and models. Given an epistemic plausibility frame S, an *S-proposition* is any subset $P \subseteq S$. Intuitively, we say that *a state s satisfies the proposition P* if $s \in P$. Observe that a plausibility frame is just a special case of a

relational frame (or *Kripke frame*). So, as it is standard for Kripke frames in general, we can define a *plausibility model* to be a structure $\mathbf{S} = (S, \leq, \|\cdot\|)$, consisting of a plausibility frame $(S, \leq)$ together with a valuation map $\|\cdot\| : \Phi \to \mathcal{P}(S)$, mapping every element of a given set Φ of "atomic sentences" into S-propositions.

Interpretation. The elements of S will represent the *possible states* (or "possible worlds") of a system. The atomic sentences $p \in \Phi$ represent *"ontic" (non-doxastic) facts*, that might hold or not in a given state. The valuation tells us which facts hold at which worlds. Finally, the plausibility relations $\leq$ capture the agent's *(conditional) beliefs about the state* of the system; if e.g., the agent was given the information that the state of the system is either s or t, she would believe that the system was in the *most plausible* of the two. So, if $s < t$, the agent would believe the real state was s; if $t < s$, she would believe it was t; otherwise (if $s \cong t$), the agent would be indifferent between the two alternatives: she will not be able to decide to believe any one alternative rather than the other.

Propositional operators, Kripke modalities. For every model $\mathbf{S}$, we have the usual Boolean operations with S-propositions

$$P \wedge Q := P \cap Q, \quad P \vee Q := P \cup Q,$$

$$\neg P := S \setminus P, \quad P \to Q := \neg P \vee Q,$$

as well as Boolean constants $\top_S := S$ and $\bot_S := \emptyset$. Obviously, one may also introduce *infinitary* conjunctions and disjunctions. In addition, any binary relation $R \subseteq S \times S$ on S gives rise to a *Kripke modality* $[R] : \mathcal{P}(S) \to \mathcal{P}(S)$, defined by

$$[R]Q := \{s \in S : \forall t\, (sRt \Rightarrow t \in Q)\}.$$

Accessibility relations for belief, conditional belief and knowledge. To talk about beliefs, we introduce a *doxastic accessibility relation* $\to$, given by

$$s \to t \ \text{ iff } \ t \in \mathrm{Min}_{\leq} S\,.$$

We read this as saying that: when the actual state is s, the agent believes that *any* of the states t with $s \to t$ *may be* the actual state. This matches the above interpretation of the preorder: the states believed to be possible are the minimal (i.e., "most plausible") ones.

In order to talk about *conditional beliefs*, we can similarly define a *conditional doxastic accessibility relation* for each S-proposition $P \subseteq S$:

$$s \to^P t \ \text{ iff } \ t \in \mathrm{Min}_{\leq} P\,.$$

We read this as saying that: when the actual state is s, if the agent is given the information (that) P (is true at the actual state), then she believes that *any* of the states t with $s \to t$ *may be* the actual state.

Finally, to talk about knowledge, we introduce a relation of *epistemic possibility* (or "indistinguishability") $\sim$. Essentially, this is just the universal relation:

$$s \sim t \text{ iff } s, t \in S.$$

So, in single-agent models, *all* the states in S are assumed to be "epistemically possible": the only thing *known* with absolute certainty about the current state is that it belongs to S. This is natural, in the context of a single agent: the states known to be impossible are *irrelevant* from the point of doxastic-epistemic logic, so they can simply be excluded from our model S. (As seen below, this cannot be done in the case of multiple agents!)

Knowledge and (conditional) belief. We define *knowledge* and *(conditional) belief* as the Kripke modalities for the epistemic and (conditional) doxastic accessibility relations:

$$KP := [\sim]P,$$

$$BP := [\rightarrow]P,$$

$$B^Q P := [\rightarrow^Q]P.$$

We read KP as saying that the (implicit) agent *knows* P. This is "knowledge" in the strong Leibnizian sense of "truth in all possible worlds". We similarly read BP as "P is believed" and $B^Q P$ as "P is believed given (or conditional on) Q". As for *conditional belief* statements $s \in B^Q P$, we interpret them in the following way: if the actual state is s, then after coming to believe that Q is the case (at this actual state), the agent will believe that P *was* the case (at the same actual state, *before* his change of belief). In other words, conditional beliefs B^Q give descriptions of the agent's *plan* (or *commitments*) about what he will believe about the current state after receiving new (believable) information. To quote Johan van Benthem in (2007), conditional beliefs are *"static pre-encodings" of the agent's potential belief changes* in the face of new information.

Discussion on interpretation. Observe that our interpretation of the plausibility relations is *qualitative*, in terms of *conditional beliefs* rather than "degrees of belief": there is no scale of beliefs here, allowing for "intermediary" stages between believing and not believing. Instead, all these beliefs are equally "firm" (*though conditional*): given the condition, something is either believed or not. To repeat, writing $s < t$ is for us just a way to say that: if *given* the information that the state of the system is either s or t, the agent would believe it to be s. So plausibility relations are special cases of conditional belief. This interpretation is based on the following (easily verifiable) equivalence:

$$s < t \text{ iff } s \in B^{\{s,t\}}\{s\} \text{ iff } t \in B^{\{s,t\}}\{s\}.$$

There is nothing quantitative here, no need for us to refer in any way to the "strength" of this agent's belief: though she might have beliefs of unequal strengths, we are not interested here in modeling this quantitative aspect. Instead, we give the agent some information about a state of a virtual system (that it is either s or t) and we ask her a *yes-or-no question* ("Do you believe that virtual state to be s ?"); we write $s < t$ iff the agent's answer is "yes". This is a firm answer, so it expresses a firm belief. "Firm" does not imply "un-revisable" though: if later we reveal to the agent that the state in question was in fact t, she should be able to accept this new information; after all, the agent should be introspective enough to realize that her belief, however firm, was just a belief.

One possible objection against this qualitative interpretation is that our postulate that $\leq$ is a well-preorder (and so in particular a connected pre-order) introduces a hidden "quantitative" feature; indeed, any such preorder can be equivalently described using a plausibility map as in e.g., Spohn (1988), assigning ordinals to states. Our answer is that, first, the specific ordinals will not play any role in our definition of a dynamic belief update; and second, all our postulates can be given a justification in purely qualitative terms, using conditional beliefs. The transitivity condition for $\leq$ is just a *consistency* requirement imposed on a rational agent's conditional beliefs. And the existence of minimal elements in any non-empty subset is simply the natural extension of the above setting to *general* conditional beliefs, not only conditions involving two states: more specifically, for any possible condition $P \subseteq S$ about a system S, the S-proposition $\mathrm{Min}_{\leq} P$ is simply a way to encode everything that the agent would believe about the current state of the system, if she was given the information that the state satisfied condition P.

Note on other models in the literature. Our models are the same as Board's "belief revision structures" (Board 2002), i.e., nothing but "Spohn models" as in Spohn (1988), but with a purely relational description. Spohn models are usually described in terms of a map assigning ordinals to states. But giving such a map is equivalent to introducing a well pre-order $\leq$ on states, and it is easy to see that all the relevant information is captured by this order.

Our conditions on the preorder $\leq$ can also be seen as a *semantic analogue* of Grove's conditions for the (relational version of his) models in Grove (1988). The standard formulation of Grove models is in terms of a "system of spheres" (weakening Lewis' similar notion), but it is equivalent (as proved in Grove 1988) to a relational formulation. Grove's postulates are still *syntax-dependent*, e.g., existence of minimal elements is required only for subsets that are *definable* in his language: this is the so-called "smoothness" condition, which is weaker than our "well-preordered" condition. We prefer a purely semantic condition, independent of the choice of a language, both for reasons of elegance and simplicity and because we want to be able to consider more than one language for the same structure.[7]

[7] Imposing syntactic-dependent conditions in the very definition of a class of structures makes the definition meaningful only for one language; or else, the meaning of what, say, a plausibility model is won't be *robust*: it will change whenever one wants to extend the logic, by adding a few more

So, following Board (2002) and Stalnaker (2006) and others, we adopt the natural semantic analogue of Grove's condition, simply requiring that *every* subset has minimal elements: this will allow our conditional operators to be well-defined on sentences of *any* extension of our logical language.

Note that the minimality condition implies, by itself, that the relation $\leq$ is both *reflexive* (i.e., $s \leq s$ for all $s \in S$) and *connected*[8] (i.e., either $s \leq t$ or $t \leq s$, for all $s, t \in S$). In fact, a "well-preorder" is the same as a *connected, transitive, well-founded*[9] *relation*, which is the setting proposed in Board (2002) for a logic of conditional beliefs equivalent to our logic CDL below. Note also that, when the set S is *finite*, a well-preorder is nothing but a *connected preorder*. This shows that our notion of frame subsumes, not only Grove's setting, but also some of the other settings proposed for conditionalization.

Multi-agent Plausibility Models

In the multi-agent case, *we cannot exclude from the model the states that are known to be impossible* by some agent a: they may still be considered possible by a second agent b. Moreover, they might still be relevant for a's beliefs/knowledge about what b believes or knows. So, in order to define an agent's knowledge, we cannot simply quantify over *all* states, as we did above: instead, we need to consider, as usually done in the Kripke-model semantics of knowledge, only the "possible" states, i.e., the ones that are *indistinguishable* from the real state, as far as a given agent is concerned. It is thus natural, in the multi-agent context, to explicitly specify the agents' epistemic indistinguishability relations $\sim_a$ (labeled with the agents' names) as part of the basic structure, in addition to the plausibility relations $\leq_a$. Taking this natural step, we obtain *epistemic plausibility frames* $(S, \sim_a, \leq_a)$. As in the case of a single agent, specifying epistemic relations turns out to be *superfluous*: the relations $\sim_a$ can be recovered from the relations $\leq_a$. Hence, we will simplify the above structures, obtaining the equivalent setting of *multi-agent plausibility frames* $(S, \leq_a)$.

Before going on to define these notions, observe that it doesn't make sense anymore to require the plausibility relations $\leq_a$ to be connected (and even less sense to require them to be well-preordered): if two states s, t are *distinguishable* by an agent a, i.e., $s \not\sim_a t$, then a will never consider both of them as epistemically possible in the same time. If she was given the information that the real state is either s or t, agent a will immediately *know* which of the two: if the real state was s, she would be able to distinguish this state from t, and would thus know the state

operators. This is very undesirable, since then one cannot compare the expressivity of different logics on the same class of models.

[8] In the Economics literature, connectedness is called "completeness", see e.g., Board (2002).

[9] I.e., there exists no infinite descending chain $s_0 > s_1 > \cdots$.

was s; similarly, if the real state was t, she would know it to be t. Her beliefs will play no role in this, and it would be meaningless to ask her which of the two states is more plausible to her. So only the states in the same $\sim_a$-equivalence class could, and should, be $\leq_a$-comparable; i.e., $s \leq_a t$ *implies* $s \sim_a t$, and the restriction of $\leq_a$ to each $\sim_a$-equivalence class is connected. Extending the same argument to arbitrary conditional beliefs, we can see that *the restriction of* $\leq_a$ *to each* $\sim_a$*-equivalence class must be well-preordered.*

Epistemic plausibility frames. Let $\mathcal{A}$ be a finite set of labels, called *agents*. An *epistemic plausibility frame* over $\mathcal{A}$ (EPF, for short) is a structure $\mathbf{S} = (S, \sim_a, \leq_a)_{a\in\mathcal{A}}$, consisting of a set S of "states", endowed with a family of equivalence relations $\sim_a$, called *epistemic indistinguishability relations*, and a family of *plausibility relations* $\leq_a$, both labeled by "agents" and assumed to satisfy two conditions: (1) $\leq_a$-comparable states are $\sim_a$-indistinguishable (i.e., $s \leq_a t$ implies $s \sim_a t$); (2) the restriction of each plausibility relation $\leq_a$ to each $\sim_a$-equivalence class is a well-preorder. As before, we use the notation $\mathrm{Min}_{\leq_a} P$ for the set of $\leq_a$-minimal elements of P. We write $s <_a t$ iff $s \leq_a t$ but $t \not\leq_a s$ (the *"strict" plausibility relation*), and write $s \cong_a t$ iff both $s \leq_a t$ and $t \leq_a s$ (the *"equi-plausibility" relation*). The notion of *epistemic plausibility models* (EPM, for short) is defined in the same way as the plausibility models in the previous section.

Epistemic plausibility models. We define a *(multi-agent) epistemic plausibility model* (EPM, for short) as a multi-agent EPF together with a valuation over it (the same way that single-agent plausibility models were defined in the previous section).

It is easy to see that our definition of EPFs includes superfluous information: in an EPF, the knowledge relation $\sim_a$ can be recovered from the plausibility relation $\leq_a$, via the following rule:

$$s \sim_a t \text{ iff either } s \leq_a t \text{ or } t \leq_a s\,.$$

In other words, two states are indistinguishable for a iff they are *comparable* (with respect to $\leq_a$).

So, in fact, one could present epistemic plausibility frames simply as *multi-agent plausibility frames*. To give this alternative presentation, we use, for *any preorder relation* $\leq$, the notation $\sim$ for the associated *comparability relation*

$$\sim \;:= \leq \cup \geq$$

(where $\geq$ is the converse of $\leq$). A *comparability class* is a set of the form $\{t : s \leq t \text{ or } t \leq s\}$, for some state s. A relation $\leq$ is called *locally well-preordered* if it is a preorder such that its restriction to each comparability class is well-preordered. Note that, when the underlying set S is *finite*, a locally well-preordered relation is nothing but a *locally connected preorder*: a preorder whose restrictions to any comparability class are connected. More generally, *a locally well-preordered relation is the same as a locally connected and well-founded preorder.*

Multi-agent plausibility frames. A *multi-agent plausibility frame* (MPF, for short) is a structure $(S, \leq_a)_{a\in\mathcal{A}}$, consisting of a set of states S together with a family of locally well-preordered relations $\leq_a$, one for each agent $a \in \mathcal{A}$. Oliver Board (2002) calls multi-agent plausibility frames "belief revision structures". A *multi-agent plausibility model* (MPM, for short) is an MPF together with a valuation map.

Bijective correspondence between EPFs and MPFs. *Every MPF can be canonically mapped into an EPF*, obtained by defining epistemic indistinguishability via the above rule ($\sim_a := \leq_a \cup \geq_a$). Conversely, every EPF gives rise to an MPF, via the map that "forgets" the indistinguishability structure. It is easy to see that these two maps are the inverse of each other. Consequently, from now on we *identify MPFs and EPFs, and similarly identify MPMs and EPMs*; e.g., we can talk about "knowledge", "(conditional) belief" etc. in an MPM, defined in terms of the associated EPM.

So from now on we identify the two classes of models, via the above canonical bijection, and talk about "plausibility models" in general. One can also see how this approach relates to another widely adopted definition for conditional beliefs; in Board (2002), van Ditmarsch (2005), and van Benthem (2007), this definition involves the assumption of a *"local plausibility" relation at a given state* $s \leq_a^w t$, to be read as: "at state w, agent a considers state s at least as plausible as state t ". Given such a relation, the conditional belief operator is usually defined in terms that are equivalent to putting $s \to_a^P t$ iff $t \in \mathrm{Min}_{\leq_a^s} P$. One could easily restate our above definition in this form, by taking:

$$s \leq_a^w t \quad \text{iff} \quad \text{either } w \not\sim_a t \text{ or } s \leq_a t.$$

The converse problem is studied in Board (2002), where it is shown that, if full introspection is assumed, then one can recover "uniform" plausibility relations $\leq_a$ from the relations $\leq_a^w$.

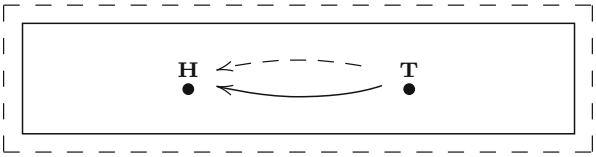
Information cell. The equivalence relation $\sim_a$ induces a partition of the state space S, called *agent a's information partition*. We denote by $s(a)$ the *information cell* of s in a's partition, i.e., the $\sim_a$-equivalence class of s:

$$s(a) := \{t \in S : s \sim_a t\}.$$

The information cell $s(a)$ captures *all the knowledge possessed by the agent* at state s: when the actual state of the system is s, then agent a knows only the state's equivalence class $s(a)$.

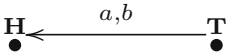
Example 1. Alice and Bob play a game, in which an anonymous referee puts a coin on the table, lying face up but in such a way that the face is covered (so Alice and Bob cannot see it). Based on previous experience, (it is common knowledge that) Alice and Bob believe that the upper face is Heads (since e.g., they noticed that

the referee had a strong preference for Heads). And in fact, they're right: the coin lies Heads up. Neglecting the anonymous referee, the EPM for this example is the following model **S**:

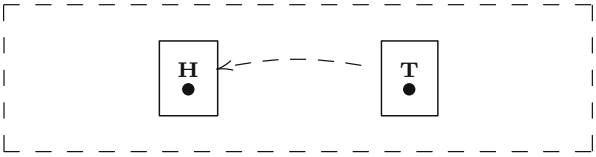


Here, the arrows represent *converse plausibility relations* $\geq$ *between distinct states only* (going from less plausible to more plausible states): since these are always reflexive, we choose to *skip all the loops* for convenience. The squares represent the *information cells* for the two agents. Instead of labels, we use *dashed arrows and squares for Alice*, while using *continuous arrows and squares for Bob*. In this picture, the actual state of the system is the state s on the left (in which **H** is true). Henceforth, in our other examples, we will refer to this particular plausibility model as **S**.

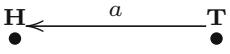
By deleting the squares, we obtain a representation of the corresponding MPM, also denoted by **S** (where we now use labels for agents instead of different types of lines):



Example 2. In front of Alice, the referee shows the face of the coin to Bob, but Alice cannot see the face. The EPM is now the following model **W**:



while the MPM is



Since Bob now knows the state of the coin, his local plausibility relation consists only of loops, and hence we have no arrows for Bob in this diagrammatic representation.

(Conditional) doxastic appearance and (conditional) doxastic accessibility. As in the previous section, we can define a doxastic and epistemic accessibility relations, except that now we have to select, for each state s, the most plausible states in its information cell $s(a)$ (instead of the most plausible in S). For this, it is convenient to introduce some notation and terminology: the *doxastic appearance* of state s to agent a is the set

$$s_a := \mathrm{Min}_{\leq_a} s(a)$$

of the "most plausible" states that are consistent with the agent's knowledge at state s. The doxastic appearance of s captures *the way state s appears to the agent*, or (in the language of Belief Revision) *the agent's current "theory" about the world s*. We can extend this to capture *conditional beliefs* (in full generality), by associating to each S-proposition $P \subseteq S$ and each state $s \in S$ the *conditional doxastic appearance* s_a^P *of state s to agent a, given (information) P*. This can be defined as the S-proposition

$$s_a^P := \mathrm{Min}_{\leq_a} s(a) \cap P$$

given by the set of all $\leq_a$-minimal states of $s(a) \cap P$: these are the "most plausible" states satisfying P that are consistent with the agent's knowledge at state s. The conditional appearance S_a^P gives *the agent's revised theory (after learning P) about the world s*. We can put these in a relational form, by defining *doxastic accessibility relations* $\rightarrow_a$, $\rightarrow_a^P$, as follows:

$$s \rightarrow_a t \text{ iff } t \in s_a\,,$$
$$s \rightarrow_a^P t \text{ iff } t \in s_a^P\,.$$

Knowledge and (conditional) belief. As before, we define the *knowledge* and *(conditional) belief* operators for an agent a as the Kripke modalities for a's epistemic and (conditional) doxastic accessibility relations:

$$K_a P := [\sim_a]P = \{s \in S : s(a) \subseteq P\}\,,$$
$$B_a P := [\rightarrow_a]P = \{s \in S : s_a \subseteq P\}\,,$$
$$B_a^Q P := [\rightarrow_a^Q]P = \{s \in S : s_a^Q \subseteq P\}.$$

We also need a notation for the *dual of the K modality* ("epistemic possibility"):

$$\tilde{K}_a P := \neg K_a \neg P.$$

Doxastic propositions. Until now, our notion of proposition is "local", being specific to a given model: we only have "S-propositions" for each model S. As long as the model is fixed, this notion is enough for interpreting sentences over the

given model. But, since later we will proceed to study systematic *changes* of models (when dealing with *dynamic* belief revision), we need a notion of proposition that is not confined to one model, but makes sense on *all* models:

A *doxastic proposition* is a map $\mathbf{P}$ assigning to each plausibility model $\mathbf{S}$ some S-proposition $\mathbf{P_S} \subseteq S$. We write $s \models_{\mathbf{S}} \mathbf{P}$, and say that the proposition $\mathbf{P}$ is true at $s \in \mathbf{S}$, iff $s \in (\mathbf{P})_{\mathbf{S}}$. We skip the subscript and write $s \models \mathbf{P}$ when the model is understood.

We denote by Prop the family of all doxastic propositions. All the Boolean operations on S-propositions as sets can be *lifted* pointwise to operations on Prop: in particular, we have the "always true" $\top$ and "always false" $\bot$ propositions, given by $(\bot)_{\mathbf{S}} := \emptyset, (\top)_{\mathbf{S}} := S$, negation $(\neg\mathbf{P})_{\mathbf{S}} := S \setminus \mathbf{P_S}$, conjunction $(\mathbf{P} \wedge \mathbf{Q})_{\mathbf{S}} := \mathbf{P_S} \cap \mathbf{Q_S}$, disjunction $(\mathbf{P} \vee \mathbf{Q})_{\mathbf{S}} := \mathbf{P_S} \cup \mathbf{Q_S}$ and all the other standard Boolean operators, including *infinitary* conjunctions and disjunctions. Similarly, we can define pointwise the *epistemic and (conditional) doxastic modalities*: $(K_a\mathbf{P})_{\mathbf{S}} := K_a\mathbf{P_S}$, $(B_a\mathbf{P})_{\mathbf{S}} := B_a\mathbf{P_S}$, $(B_a^{\mathbf{Q}}\mathbf{P})_{\mathbf{S}} := B_a^{\mathbf{Q_S}}\mathbf{P_S}$. It is easy to check that we have: $B_a\mathbf{P} = B_a^{\top}\mathbf{P}$. Finally, the relation of *entailment* $\mathbf{P} \models \mathbf{Q}$ between doxastic propositions is given pointwise by inclusion: $\mathbf{P} \models \mathbf{Q}$ iff $\mathbf{P_S} \subseteq \mathbf{Q_S}$ for all $\mathbf{S}$.

Safe Belief and the Defeasibility Theory of Knowledge

Ever since Plato's *identification of knowledge with "true justified (or justifiable) belief"* was shattered by Gettier's celebrated counterexamples (Gettier 1963), philosophers have been looking for the "missing ingredient" in the Platonic equation. Various authors identify this missing ingredient as "robustness" (Hintikka 1962), "indefeasibility" (Klein 1971; Lehrer 1990; Lehrer and Paxson 1969; Stalnaker 2006) or "stability" (Rott 2004). According to this *defeasibility theory of knowledge* (or "stability theory", as formulated by Rott), a belief counts as "knowledge" if it is *stable under belief revision with any new evidence*: "if a person has knowledge, than that person's justification must be sufficiently strong that it is not capable of being defeated by evidence that he does not possess" (Pappas and Swain 1978).

One of the problems is interpreting what "evidence" means in this context. There are at least two natural interpretations, each giving us a concept of "knowledge". The first, and the most common,[10] interpretation is to take it as meaning "any *true* information". The resulting notion of "knowledge" was formalized by Stalnaker in (2006), and defined there as follows: "an agent knows that φ if and only if φ is true, she believes that φ, and she continues to believe φ if any *true* information is received". This concept differs from the usual notion of knowledge ("Aumann knowledge") in Computer Science and Economics, by the fact that it does not satisfy the laws of the modal system S5 (in fact, negative introspection fails); Stalnaker

[10]This interpretation is the one virtually adopted by all the proponents of the defeasibility theory, from Lehrer to Stalnaker.

shows that the complete modal logic of this modality is the modal system S4.3. As we'll see, this notion ("Stalnaker knowledge") corresponds to what we call "safe belief" $\Box P$. On the other hand, another natural interpretation, considered by at least one author Rott (2004), takes "evidence" to mean "*any* proposition", i.e., to include possible *misinformation*: "real knowledge" should be robust even in the face of false evidence. As shown below, this corresponds to our "knowledge" modality KP, which could be called "absolutely unrevisable belief". This is a partition-based concept of knowledge, identifiable with "Aumann knowledge" and satisfying all the laws of S5. In other words, this last interpretation provides a perfectly decent "defeasibility" defense of S5 and of negative introspection!

In this paper, we adopt the pragmatic point of view of the formal logician: instead of debating which of the two types of "knowledge" is the real one, we simply formalize both notions in a common setting, compare them, axiomatize the logic obtained by combining them and use their joint strength to express interesting properties. Indeed, as shown below, conditional beliefs can be *defined* in terms of knowledge *only* if we combine both the above-mentioned types of "knowledge".

Knowledge as unrevisable belief. Observe that, for all propositions $\mathbf{P}$, we have

$$K_a\mathbf{Q} = \bigwedge_{\mathbf{P}} B_a^{\mathbf{P}}\mathbf{Q}$$

(where the conjunction ranges over *all* doxastic propositions), or equivalently, we have for every state s in every model $\mathbf{S}$:

$$s \models K_a\mathbf{Q} \quad \text{iff} \quad s \models B_a^{\mathbf{P}}\mathbf{Q} \text{ for all } \mathbf{P}. \tag{39.1}$$

This gives a characterization of *knowledge as "absolute" belief, invariant under any belief revision*: a given belief is "known" iff it cannot be revised, i.e., it would be still believed in any condition.[11] Observe that this resembles the defeasibility analysis of knowledge, but only if we adopt the *second interpretation* mentioned above (taking "evidence" to include misinformation). Thus, our "knowledge" is more robust than Stalnaker's: it resists any belief revision, not capable of being defeated by *any* evidence (including false evidence). This is a very "strong" notion of knowledge (implying "absolute certainty" and full introspection), which seems to us to fit better with the standard usage of the term in Computer Science literature. Also, unlike the one in Stalnaker (2006), our notion of knowledge *is negatively introspective*.

[11]This of course assumes agents to be "rational" in a sense that excludes "fundamentalist" or "dogmatic" beliefs, i.e., beliefs in unknown propositions but refusing any revision, even when contradicted by facts. But this "rationality" assumption is already built in our plausibility models, which satisfy an epistemically friendly version of the standard AGM postulates of rational belief revision. See Baltag and Smets (2006a) for details.

Another identity[12] that can be easily checked is:

$$K_a\mathbf{Q} = B_a^{\neg\mathbf{Q}}\mathbf{Q} = B_a^{\neg\mathbf{Q}}\bot \tag{39.2}$$

(where $\bot$ is the "always false" proposition). This captures in a different way the "absolute un-revisability" of knowledge: something is "known" if it is believed even if conditionalizing our belief with its negation. In other words, this simply expresses the *impossibility* of accepting its negation as evidence (since such a revision would lead to an inconsistent belief).

Safe belief. To capture "Stalnaker knowledge", we introduce the Kripke modality $\Box_a$ associated to the converse $\geq_a$ of the plausibility relation, going from any state s to all the states that are "at least as plausible" as s. For S-propositions $P \subseteq S$ over any given model $\mathbf{S}$, we put

$$\Box_a P := [\geq_a]P = \{s \in S : t \in P \text{ for all } t \leq_a s\},$$

and this induces pointwise an operator $\Box_a\mathbf{P}$ on doxastic propositions. We read $s \models \Box_a\mathbf{P}$ as saying that: *at state s, agent a's belief in* $\mathbf{P}$ *is safe*; or at *state s, a safely believes that* $\mathbf{P}$. We will explain this reading below, but first observe that: $\Box_a$ is an $S4$-modality (since $\geq_a$ is reflexive and transitive), but not necessarily $S5$; i.e., *safe beliefs are truthful* ($\Box_a\mathbf{P} \models \mathbf{P}$) and *positively introspective* ($\Box_a\mathbf{P} \models \Box_a\Box_a\mathbf{P}$), but not necessarily negatively introspective: in general, $\neg\Box_a\mathbf{P} \not\models \Box_a\neg\Box_a\mathbf{P}$.

Relations between knowledge, safe belief and conditional belief. First, *knowledge entails safe belief*

$$K_a\mathbf{P} \models \Box_a\mathbf{P}\,,$$

and *safe belief entails belief*

$$\Box_a\mathbf{P} \models B_a\mathbf{P}.$$

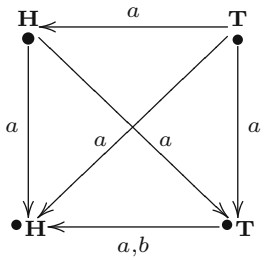
The last observation can be strengthened to characterize safe belief in a similar way to the above characterization (39.1) of knowledge (as belief invariant under any revision): *safe beliefs are precisely the beliefs which are persistent under revision with any true information*. Formally, this says that, for every state s in every model $\mathbf{S}$, we have

$$s \models \Box_a\mathbf{Q} \quad \text{iff:} \quad s \models B_a^{\mathbf{P}}\mathbf{Q} \text{ for every } \mathbf{P} \text{ such that } s \models \mathbf{P} \tag{39.3}$$

[12]This identity corresponds to the definition of "necessity" in Stalnaker (1968) in terms of doxastic conditionals.

We can thus see that *safe belief coincides indeed with Stalnaker's notion of "knowledge"*, given by the first interpretation ("evidence as true information") of the defeasibility theory. As mentioned above, we prefer to keep the name "knowledge" for the strong notion (which gives absolute certainty), and call this weaker notion "safe belief": indeed, these are beliefs that are "safe" to hold, in the sense that no future learning of truthful information will force us to revise them.

Example 3 (Dangerous Knowledge). This starts with the situation in Example 1 (when none of the two agents has yet seen the face of the coin). Alice has to get out of the room for a minute, which creates an opportunity for Bob to quickly raise the cover in her absence and take a peek at the coin. He does that, and so he sees that the coin is Heads up. After Alice returns, she obviously doesn't know whether or not Bob took a peek at the coin, but she believes he didn't do it: taking a peek is against the rules of the game, and so she trusts Bob not to do that. The model is now rather complicated, so we only represent the MPM:



Let us call this model $\mathbf{S}'$. The actual state s'_1 is the one in the upper left corner, in which Bob took a peek and saw the coin Heads up, while the state t'_1 in the upper right corner represents the other possibility, in which Bob saw the coin lying Tails up. The two lower states s'_2 and t'_2 represent the case in which Bob *didn't take a peek*. Observe that the above drawing includes the (natural) assumption that Alice keeps her previous belief that the coin lies Heads up (since there is no reason for her to change her mind). Moreover, we assumed that she will keep this belief even if she'd be told that Bob took a peek: this is captured by the a-arrow from t'_1 to s'_1. This seems natural: Bob's taking a peek doesn't change the upper face of the coin, so it shouldn't affect Alice's prior belief about the coin.

In both Examples 1 and 3 above, Alice holds a *true belief* (at the real state) that the coin lies Heads up: the actual state satisfies $B_a\mathbf{H}$. In both cases, this true belief is *not knowledge* (since Alice doesn't know the upper face), but nevertheless in Example 1, this belief is *safe* (although it is *not known by the agent to be safe*): no additional truthful information (about the real state s) can force her to revise this belief. (To see this, observe that any *new* truthful information would reveal to Alice the real state s, thus confirming her belief that Heads is up.) So in the model $\mathbf{S}$ from Example 1, we have $s \models \Box_a\mathbf{H}$ (where s is the actual state). In contrast, in Example 2, Alice's belief (that the coin is Heads up), though true, is *not safe*. There is some piece

of correct information (about the real state s'_1) which, if learned by Alice, would make her change this belief: we can represent this piece of correct information as the doxastic proposition $\mathbf{H} \to \mathbf{K_b H}$. It is easy to see that the actual state s'_1 of the model $\mathbf{S}'$ satisfies the proposition $B_a^{\mathbf{H} \to \mathbf{K_b H}} \mathbf{T}$ (since $(\mathbf{H} \to K_b \mathbf{H})_{\mathbf{S}'} = \{s'_1, t'_1, t'_2\}$ and the minimal state in the set $s'_1(a) \cap \{s'_1, s'_1, t'_2\} = \{s'_1, t'_1, t'_2\}$ is t'_2, which satisfies $\mathbf{T}$.) So, if given this information, Alice would come to wrongly believe that the coin is Tails up! This is an example of a *dangerous truth*: a true information whose learning can lead to wrong beliefs.

Observe that *an agent's belief can be safe without him necessarily knowing this* (in the "strong" sense of knowledge given by K): "safety" (similarly to "truth") is an *external* property of the agent's beliefs, that can be ascertained only by comparing his belief-revision system with reality. Indeed, *the only way* for an agent to *know a belief to be safe* is to actually *know it to be truthful*, i.e., to have actual *knowledge* (not just a belief) of its truth. This is captured by the valid identity

$$K_a \Box_a \mathbf{P} = K_a \mathbf{P}. \tag{39.4}$$

In other words: *knowing that something is safe to believe is the same as just knowing it to be true*. In fact, *all beliefs held by an agent "appear safe" to him*: in order to believe them, he has to believe that they are safe. This is expressed by the valid identity

$$B_a \Box_a \mathbf{P} = B_a \mathbf{P} \tag{39.5}$$

saying that: *believing that something is safe to believe is the same as just believing it*. Contrast this with the situation concerning "knowledge": in our logic (as in most standard doxastic-epistemic logics), we have the identity

$$B_a K_a \mathbf{P} = K_a \mathbf{P}. \tag{39.6}$$

So *believing that something is known is the same as knowing it*!

The Puzzle of the Perfect Believer. The last identity is well-known and has been considered "paradoxical" by many authors. In fact, the so-called "Paradox of the Perfect Believer" in Gochet and Gribomont (2006), Voorbraak (1993), Hoek (1993), Meyer and Hoek (1995), Williamson (2001), and Friedmann and Halpern (1994) is based on it. For a "strong" notion of belief as the one we have here ("belief" = belief with certainty), it seems reasonable to assume the following "axiom":

$$B_a \varphi \to B_a K_a \varphi. \tag{?}$$

Putting this together with (39.6) above, we get a paradoxical conclusion:

$$B_a \varphi \to K_a \varphi. \tag{?!}$$

So this leads to a triviality result: *knowledge and belief collapse to the same thing*, and *all beliefs are always true*! One solution to the "paradox" is to reject (?), as an (intuitive but) *wrong* "axiom". In contrast, various authors Friedmann and Halpern (1994); Hoek (1993); Voorbraak (1993); Williamson (2001) accept (?) and propose other solutions, e.g., giving up the principle of "negative introspection" for knowledge.

Our solution to the paradox, as embodied in the contrasting identities (39.5) and (39.6), combines the advantages of both solutions above: the "axiom" (?) is *correct if we interpret "knowledge" as safe belief* $\Box_a$, since then (?) becomes equivalent to identity (39.5) above; but then *negative introspection fails for this interpretation*! On the other hand, if we interpret "knowledge" as our K_a-modality then negative introspection holds; but then *the above "axiom" (?) fails*, and on the contrary we have the identity (39.6).

So, in our view, *the paradox of the perfect believer arises from the conflation of two different notions of "knowledge":* "Aumann" (partition-based) knowledge and "Stalnaker" knowledge (i.e., safe belief).

(Conditional) beliefs in terms of "knowledge" notions. An important observation is that *one can characterize/define (conditional) beliefs only in terms of our two "knowledge" concepts (K and* $\Box$*)*: For simple beliefs, we have

$$B_a\mathbf{P} \;=\; \tilde{K}_a\Box_a\mathbf{P} \;=\; \Diamond_a\Box_a\mathbf{P}\,,$$

recalling that $\tilde{K}_a\mathbf{P} = \neg K_a\neg\mathbf{P}$ is the Diamond modality for K_a, and $\Diamond_a\mathbf{P} = \neg\Box_a\neg\mathbf{P}$ is the Diamond for $\Box_a$.

The equivalence $B_a\mathbf{P} \;=\; \Diamond_a\Box_a\mathbf{P}$ has recently been observed by Stalnaker in (2006), who took it as the basis of a philosophical analysis of "belief" in terms of "defeasible knowledge" (i.e., safe belief). Unfortunately, this analysis does not apply to conditional belief: one can easily see that *conditional belief cannot be defined in terms of safe belief only*! However, one *can generalize* the identity $B_a\mathbf{P} = \tilde{K}_a\Box_a\mathbf{P}$ above, defining conditional belief in terms of *both our "knowledge" concepts*:

$$B_a^{\mathbf{P}}\mathbf{Q} \;=\; \tilde{K}_a\mathbf{P} \to \tilde{K}_a(\mathbf{P} \wedge \Box_a(\mathbf{P} \to \mathbf{Q}))\,. \tag{39.7}$$

Other Modalities and Doxastic Attitudes

From a modal logic perspective, it is natural to introduce the Kripke modalities $[>_a]$ and $[\cong_a]$ for the other important relations (strict plausibility and equiplausibility): For S-propositions $P \subseteq S$ over a given model $\mathbf{S}$, we put

$$[>_a]P \;:=\; \{s \in S : t \in P \text{ for all } t <_a s\},$$
$$[\cong_a]P \;:=\; \{s \in S : t \in P \text{ for all } t \cong_a s\},$$

and as before these pointwise induce corresponding operators on Prop. The intuitive meaning of these operators is not very clear, but they can be used to define other interesting modalities, capturing various "doxastic attitudes".

Weakly safe belief. We can define a *weakly safe belief* operator $\Box^{\text{weak}}\mathbf{P}$ in terms of the strict order by putting:

$$\Box_a^{\text{weak}}\mathbf{P} \;=\; \mathbf{P} \wedge [>_a]\mathbf{P}\,.$$

Clearly, this gives us the following truth clause:

$$s \models \Box_a^{\text{weak}}\ \mathbf{P} \text{ iff: } \ s \models \mathbf{P} \text{ and } t \models \mathbf{P} \text{ for all } t < s\,.$$

But a more useful characterization is the following:

$$s \models \Box_a^{\text{weak}}\mathbf{Q} \quad \text{iff:} \quad s \models \neg B_a^{\mathbf{P}}\neg\mathbf{Q} \text{ for every } \mathbf{P} \text{ such that } s \models \mathbf{P}\,.$$

So "weakly safe beliefs" are *beliefs which (might be lost but) are never reversed (into believing the opposite) when revising with any true information*.

The unary revision operator. Using the strict plausibility modality, we can also define a unary "belief revision" modality $*_a$, which in some sense *internalizes the standard (binary) belief revision operator*, by putting:

$$*_a\mathbf{P} \;=\; \mathbf{P} \wedge [>_a]\neg\mathbf{P}\,.$$

This gives us the following truth clause:

$$s \models *_a\mathbf{P} \text{ iff } s \in s_a^{\mathbf{P}}\,.$$

It is easy to see that $*_a\mathbf{P}$ *selects from any given information cell* $s(a)$ *precisely those states that satisfy agent a's revised theory* s_a^P:

$$*_aP \cap s(a) = s_a^P\,.$$

Recall that $s_a^P \;=\; \text{Min}_{\leq_a} s(a) \cap P$ is the conditional appearance of s to a given P, representing the agent's "revised theory" (after revision with P) about s. This explains our interpretation: the proposition $*_a\mathbf{P}$ is a *complete description of the agent's P-revised "theory" about the current state*.

Another interesting identity is the following:

$$B_a^{\mathbf{P}}\mathbf{Q} \;=\; K_a(*_a\mathbf{P} \to \mathbf{Q})\,. \tag{39.8}$$

In other words: $\mathbf{Q}$ is *a conditional belief (given a condition* $\mathbf{P}$*) iff it is a known consequence of the agent's revised theory (after revision with* $\mathbf{P}$*)*.

Degrees of belief. Spohn's "degrees of belief" in Spohn (1988) were captured by Aucher (2003) and van Ditmarsch (2005) using logical operators $B_a^n\mathbf{P}$. Intuitively, 0-belief $B_a^0\mathbf{P}$ is the same as simple belief $B_a\mathbf{P}$; 1-belief $B_a^1\mathbf{P}$ means that $\mathbf{P}$ is believed conditional on learning that not all the 0-beliefs are true etc. Formally, this can be introduced e.g., by defining by induction a sequence of appearance maps s_a^n for all states s and natural numbers n:

$$s_a^0 = \mathrm{Min}_{\leq_a} s(a) \ , \ s_a^n = \mathrm{Min}_{\leq_a}\Big(s(a) \setminus \bigcup_{i<n} s_a^i\Big)$$

and defining

$$s \models B_a^n\mathbf{P} \ \text{ iff } \ t \models \mathbf{P} \text{ for all } t \in s_a^n.$$

A state s has degree of belief n if we have $s \in s_a^n$. An interesting observation is that the *finite degrees of belief* $B_a^n\mathbf{P}$ *can be defined using the unary revision operator* $*_a\mathbf{P}$ *and the knowledge operator* K_a (and, as a consequence, they can be defined using the plausibility operator $[>_a]\mathbf{P}$ and the knowledge operator). To do this, first put inductively:

$$b_a^0 := *_a\top \ , \ b_a^n := *_a\Big(\bigwedge_{m<n} \neg b_a^m\Big) \ \text{ for all } n \geq 1$$

and then put

$$B_a^n\mathbf{P} \ := \ \bigwedge_{m<n} \neg K_a(b_a^m \to \mathbf{P}) \wedge K_a(b_a^n \to \mathbf{P}) \, .$$

"Strong belief". Another important doxastic attitude can be defined in terms of knowledge and safe belief as:

$$Sb_a\mathbf{P} \ = \ B_a\mathbf{P} \wedge K_a(\mathbf{P} \to \Box_a\mathbf{P}) \, .$$

In terms of the plausibility order, it means that *all the* $\mathbf{P}$*-states in the information cell* $s(a)$ *of* s *are bellow (more plausible than) all the non-*$\mathbf{P}$ *states in* $s(a)$ (and that, moreover, *there are such* $\mathbf{P}$*-states in* $s(a)$). This notion is called "strong belief" by Battigalli and Siniscalchi (2002), while Stalnaker (1996) calls it "robust belief". Another characterization of strong belief is the following

$s \models Sb_a\mathbf{Q}$ iff:

$$s \models B_a\mathbf{Q} \text{ and } s \models B_a^{\mathbf{P}}\mathbf{Q} \text{ for every } \mathbf{P} \text{ such that } s \models \neg K_a(\mathbf{P} \to \neg\mathbf{Q}).$$

In other words: *something is strong belief if it is believed and if this belief can only be defeated by evidence (truthful or not) that is known to contradict it*. An example is the "presumption of innocence" in a trial: requiring the members of the jury to hold the accused as "innocent until proven guilty" means asking them to start the trial with a "strong belief" in innocence.

The Logic of Conditional Beliefs

The logic CDL ("conditional doxastic logic") introduced in Baltag and Smets (2006a) is a logic of conditional beliefs, equivalent to the strongest logic considered in Board (2002). The *syntax* of CDL (without common knowledge and common belief operators[13]) is:

$$\varphi \ := \ p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B_a^{\varphi}\varphi$$

while the *semantics* is given by an *interpretation map* associating to each sentence φ of CDL a doxastic proposition $\|\varphi\|$. The definition is by induction, in terms of the obvious compositional clauses (using the doxastic operators $B_a^{\mathbf{P}}\mathbf{Q}$ defined above).

In this logic, *knowledge and simple (unconditional) belief are derived operators*, defined as abbreviations by putting $K_a\varphi := B_a^{\neg\varphi}\varphi$, $B_a\varphi := B_a^{\top}\varphi$ (where $\top := \neg(p \wedge \neg p)$ is some tautological sentence).

Proof system. In addition to the rules and axioms of propositional logic, the *proof system* of CDL includes the following:

Necessitation Rule:	From $\vdash \varphi$ infer $\vdash B_a^{\psi}\varphi$.
Normality:	$\vdash B_a^{\theta}(\varphi \to \psi) \to (B_a^{\theta}\varphi \to B_a^{\theta}\psi)$
Truthfulness of Knowledge:	$\vdash K_a\varphi \to \varphi$
Persistence of Knowledge:	$\vdash K_a\varphi \to B_a^{\theta}\varphi$
Full Introspection:	$\vdash B_a^{\theta}\varphi \to K_a B_a^{\theta}\varphi$,
	$\vdash \neg B_a^{\theta}\varphi \to K_a \neg B_a^{\theta}\varphi$
Success of Belief Revision:	$\vdash B_a^{\varphi}\varphi$
Minimality of Revision:	$\vdash \neg B_a^{\varphi}\neg\psi \to (B_a^{\varphi\wedge\psi}\theta \leftrightarrow B_a^{\varphi}(\psi \to \theta))$

Proposition 4 (Completeness and Decidability). *The above system is complete for MPMs (and so also for EPMs). Moreover, it is decidable and has the finite model property.*

Proof. The proof is essentially the same as in Board (2002). It is easy to see that the proof system above is equivalent to Board's strongest logic in Board (2002) (the one that includes axiom for full introspection), and that our models are equivalent to the "full introspective" version of the semantics in Board (2002). Q.E.D.

[13]The logic in Baltag and Smets (2006a) has these operators, but for simplicity we decided to leave them aside in this presentation.

The Logic of Knowledge and Safe Belief

The problem of finding a complete axiomatization of the logic of "defeasible knowledge" (safe belief) and conditional belief was posed as an *open question* in Board (2002). We answer this question here, by extending the logic CDL above to a complete logic $K\Box$ of *knowledge and safe belief*. Since this logic can *define* conditional belief, it is in fact equivalent to the logic whose axiomatization was required in Board (2002). Solving the question posed there becomes in fact trivial, once we observe that the higher-order definition of "defeasible knowledge" in Stalnaker (2006) and Board (2002) (corresponding to our identity (39.3) above) is in fact equivalent to our simpler, first-order definition of "safe belief" as a Kripke modality.

Syntax and semantics. The *syntax* of the logic $K\Box$ is:

$$\varphi \ := \ p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_a\varphi \mid K_a\varphi$$

while the *semantics* over plausibility models is given as for CDL, by inductively defining an interpretation map from sentences to doxastic propositions, using the obvious compositional clauses. *Belief and conditional belief are derived operators here, defined as abbreviations*:

$$\begin{aligned} B_a^{\varphi}\psi \ &:= \ \tilde{K}_a\varphi \to \tilde{K}_a(\varphi \wedge \Box_a(\varphi \to \psi)) \,, \\ B_a\varphi \ &:= \ B_a^{\top}\varphi \,, \end{aligned}$$

where $\tilde{K}_a\varphi := \neg K_a\neg\varphi$ is the Diamond modality for K, and $\top = \neg(p \wedge \neg p)$ is some tautological sentence. So *the logic $K\Box$ is more expressive than CDL.*

Proof system. In addition to the rules and axioms of propositional logic, the *proof system* for the logic $K\Box$ includes the following:

- the Necessitation Rules for both K_a and $\Box_a$;
- the $S5$-axioms for K_a;
- the $S4$-axioms for $\Box_a$;
- $K_aP \to \Box_aP$;
- $K_a(P \vee \Box_aQ) \wedge K_a(Q \vee \Box_aP) \to K_aP \vee K_aQ$.

Theorem 5 (Completeness and Decidability). *The logic $K\Box$ is* (weakly) complete *with respect to MPMs (and so also with respect to EPMs). Moreover, it is* decidable *and has* the finite model property.

Proof. A *non-standard frame (model)* is a structure $(S, \geq_a, \sim_a)_a$ (together with a valuation, in the case of models) such that $\sim_a$ are equivalence relations, $\geq_a$ are preorders, $\geq_a \ \subseteq \ \sim_a$ and the restriction of $\geq_a$ to each $\sim_a$-equivalence class is connected. For a logic with two modalities, $\Box_a$ for $\geq_a$ and K_a for the relation $\sim_a$, we can use well-known results in Modal Correspondence Theory to see

that each of these semantic conditions corresponds to one of our modal axioms above. By general classical results on canonicity and modal correspondence,[14] we immediately obtain *completeness for non-standard models*. *Finite model property for these non-standard models* follows from the same general results. But every *finite* strict preorder relation $>$ is well-founded, and an MPM is nothing but a non-standard model whose strict preorders $>_a$ are well-founded. So *completeness for ("standard") MPMs* immediately follows. Then we can use Proposition 4 above to obtain *completeness for EPMs*. Finally, *decidability* follows, in the usual way, from finite model property together with *completeness* (with respect to a *finitary* proof system) and with the *decidability of model-checking on finite models*. (This last property is obvious, given the semantics.) Q.E.D.

"Dynamic" Belief Revision

The revision captured by conditional beliefs is of a *static*, purely *hypothetical*, nature. We *cannot* interpret B_a^φ as referring to the agent's revised beliefs about the situation *after revision*; if we did, then the "Success" axiom

$$\vdash B_a^\varphi \varphi$$

would *fail for higher-level beliefs*. To see this, consider a "Moore sentence"

$$\varphi := p \wedge \neg B_a p\,,$$

saying that some fact p holds but that agent a doesn't believe it. The sentence φ is consistent, so it may very well happen to be true. But agent a's beliefs about the situation after learning that φ was true *cannot* possibly include the sentence φ itself: after learning this sentence, agent *a knows p*, and so he believes p, contrary to what φ asserts. Thus, after learning φ, agent *a knows that* φ *is false now* (after the learning). This directly contradicts the Success axiom: far from believing the sentence after learning it to be true, the agent (knows, and so he correctly) believes that it has become false. There is nothing paradoxical about this: sentences may obviously change their truth values, due to our actions. Since learning the truth of a sentence is itself an action, it is perfectly consistent to have a case in which learning changes the truth value of the very sentence that is being learnt. Indeed, this is always the case with Moore sentences. Though not paradoxical, the existence of Moore sentences shows that the "Success" axiom does not correctly describe a rational agent's (higher-level) beliefs about what is the case after a new truth is being learnt.

[14] See e.g., Blackburn et al. (2001) for the general theory of modal correspondence and canonicity.

The only way to understand the "Success" axiom in the context of higher-level beliefs is to insist on the above-mentioned "static" interpretation of conditional belief operators B_a^{φ}, as expressing the agent's *revised belief* about how the state of the world *was before the revision*.

In contrast, a *belief update* is a dynamic form of belief revision, meant to capture the actual change of beliefs induced by learning: the updated belief is about the state of the world as it is *after the update*. As noticed in Gerbrandy (1999), Baltag et al. (1998), and Baltag and Moss (2004), the original model does not usually include enough states to capture all the epistemic possibilities that arise in this way. While in the previous section the models were kept unchanged during the revision, all the possibilities being already there (so that both the unconditional and the conditional beliefs *referred to the same model*), we now have to allow for belief updates that *change the original model*.

In Baltag and Moss (2004), it was argued that *epistemic events should be modeled in essentially the same way as epistemic states*, and this common setting was taken to be given by *epistemic Kripke models*. Since in this paper we enriched our state models with doxastic plausibility relations to deal with (conditional) beliefs, it is natural to follow Baltag and Moss (2004) into extending the similarity between actions and states to this setting, thus obtaining *(epistemic) action plausibility models*. The idea of such an extension was first developed in Aucher (2003) (for a different notion of plausibility model and a different notion of update product), then generalized in van Ditmarsch (2005), where many types of action plausibility models and notions of update product, that extend the so-called *Baltag-Moss-Solecki (BMS) update product* from Baltag et al. (1998) and Baltag and Moss (2004), are explored. But both these works are based on a *quantitative* interpretation of plausibility ordinals (as "degrees of belief"), and thus they define the various types of products using complex formulas of transfinite ordinal arithmetic, for which no intuitive justification is provided.

In contrast, our notion of update product is a *purely qualitative one*, based on a *simple and intuitive relational definition*: the simplest way to define a total pre-order on a Cartesian product, given total pre-orders on each of the components, is to use either the *lexicographic* or the *anti-lexicographic* order. We choose the second option, as the closest in spirit to the classical AGM theory: it gives *priority to the new, incoming information* (i.e., to "actions" in our sense).[15] We justify this choice by interpreting the action plausibility model as representing the agent's *"incoming" belief*, i.e., the *belief-updating event*, which *"performs" the update, by "acting" on the "prior" beliefs* (as given in the state plausibility model).

[15]This choice can be seen as a generalization of the so-called *"maximal-Spohn" revision*.

Action Models

An *action plausibility model*[16] (APM, for short) is a plausibility frame $(\Sigma, \leq_a)_{a\in\mathcal{A}}$ together with a *precondition map* $\text{pre} : \Sigma \to \mathsf{Prop}$, associating to each element of Σ some doxastic proposition pre_σ. We call the elements of Σ *(basic) doxastic actions* (or "events"), and we call pre_σ the *precondition* of action σ. The basic actions $\sigma \in \Sigma$ are taken to represent *deterministic belief-revising actions* of a particularly simple nature. Intuitively, the precondition defines the *domain of applicability* of action σ: it can be executed on a state s iff s satisfies its precondition. The relations $\leq_a$ give the agents' beliefs about which actions are more plausible than others.

To model *non-determinism*, we introduce the notion of epistemic program. A *doxastic program over a given action model* $\mathbf{\Sigma}$ (or $\mathbf{\Sigma}$*-program*, for short) is simply a *set* $\Gamma \subseteq \Sigma$ of doxastic actions. We can think of doxastic programs as non-deterministic actions: each of the basic actions $\gamma \in \Gamma$ is a possible "deterministic resolution" of Γ. For simplicity, when $\Gamma = \{\gamma\}$ is a singleton, we ambiguously identify the program Γ with the action γ.

Observe that $\mathbf{\Sigma}$-programs $\Gamma \subseteq \Sigma$ are formally the "dynamic analogues" of $\mathbf{S}$-propositions $P \subseteq S$. So the dynamic analogue of the conditional doxastic appearance s_a^P (representing agent a's revised theory about state s, after revision with proposition P) is the set σ_a^Γ.

Interpretation: beliefs about changes encode changes of beliefs. The name "doxastic actions" might be a bit misleading, and from a philosophical perspective Johan van Benthem's term "doxastic events" seems more appropriate. The elements of a plausibility model do not carry information about agency or intentionality and cannot represent "real" actions in all their complexity, but only the *doxastic changes* induced by these actions: each of the nodes of the graph represents a *specific kind of change of beliefs (of all the agents)*. As in Baltag and Moss (2004), we only deal here with pure "belief changes", i.e., actions that do not change the "ontic" facts of the world, but only the agents' beliefs.[17] Moreover, we think of these as *deterministic* changes: there is at most one output of applying an action to a state.[18] Intuitively, the precondition defines the *domain of applicability* of σ: this action can be executed on a state s iff s satisfies its precondition. The plausibility pre-orderings $\leq_a$ give *the agents' conditional beliefs about the current action*. But this should be interpreted as *beliefs about changes*, that *encode changes of beliefs*. In this sense, we use such

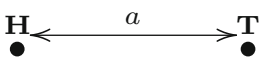
[16] Van Benthem calls this an "event model".

[17] We stress this is a minor restriction, and it is very easy to extend this setting to "ontic" actions. The only reason we stick with this restriction is that it simplifies the definitions, and that it is general enough to apply to all the actions we are interested here, and in particular to all *communication actions*.

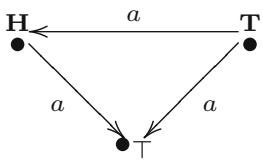
[18] As in Baltag and Moss (2004), we will be able to represent non-deterministic actions as sums (unions) of deterministic ones.

"beliefs about actions" as a way to represent doxastic changes: the information about how the agent changes her beliefs is captured by our action plausibility relations. So we read $\sigma <_a \sigma'$ as saying that: if agent a is informed that either σ or σ' is currently happening, then she cannot distinguish between the two, but she believes that σ is in fact happening. As already mentioned, doxastic programs $\Gamma \subseteq \Sigma$ represent *non-deterministic* changes of belief. Finally, for an action σ and a program Γ, the program σ_a^Γ represents *the agent's revised theory (belief) about the current action σ after "learning" that (one of the deterministic resolutions γ in) Γ is currently happening*.

Example 4 (Private "Fair-Game" Announcements). Let us consider the *action* that produced the situation represented in Example 2 above. In front of Alice, Bob looked at the coin, in such a way that (it was common knowledge that) only he saw the face. In the DEL literature, this is sometimes known as a "fair game" announcement: everybody is commonly aware that an insider (or a group of insiders) privately learns some information. It is "fair" since the outsiders are *not "deceived" in any way*: e.g., in our example, Alice knows that Bob looks at the coin (and he knows that she knows etc.). In other words, Bob's looking at the coin is not an "illegal" action, but one that obeys the (commonly agreed) "rules of the game". To make this precise, let us assume that this is happening in such a way that Alice has no strong beliefs about which of the two possible actions (Bob-seeing-Heads-up and Bob-seeing-Tails-up) is actually happening. Of course, we assumed that before this, she already believed that the coin lies Heads up, but apart from this we now assume that *the way the action (of "Bob looking") is happening gives her no indication of what face he is seeing*. We represent these actions using a two-node plausibility model $\mathbf{\Sigma}_2$ (where as in the case of state models we draw arrows for the converse plausibility relations $\geq_a$, disregarding all the loops):



Example 5 (Fully Private Announcements). Let us consider the *action* that produced the situation represented in Example 3 above. This was the action of Bob taking a peek at the coin, while Alice was away. Recall that we assumed that Alice *believed that nothing was really happening* in her absence (since she assumed Bob was playing by the rules), though obviously she *didn't know* this (that nothing was happening). In the DEL literature, this action is usually called *a fully private announcement*: Bob learns which face is up, while the outsider Alice believes nothing of the kind is happening. To represent this, we consider an action model $\mathbf{\Sigma}_3$ consisting of three "actions": the actual action σ in which Bob takes a peek and sees the coin lying Heads up; the alternative possible action ρ is the one in which Bob sees the coin lying Tails up; finally, the action τ is the one in which "nothing is really happening" (as Alice believes). The plausibility model $\mathbf{\Sigma}_3$ for this action is:



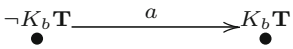
Here, the action σ is the one in the upper left corner, having precondition **H**: indeed, this can happen iff the coin is really lying Heads up; similarly, the action ρ in the upper right corner has precondition **T**, since it can only happen iff the coin is Tails up. Finally, the action τ is the lower one, having as precondition the "universally true" proposition $\top$: indeed, this action can always happen (since in it, nothing is really happening!). The plausibility relations reflect the agents' beliefs: in each case, both Bob and Charles know exactly what is happening, so their local plausibility relations are the identity (and thus we draw no arrows for them). Alice believes nothing is happening, so τ is the most plausible action for her (to which all her arrows are pointing); so she keeps her belief that **H** is the case, thus considering σ as more plausible than ρ.

Examples of doxastic programs. Consider the program $\Gamma = \{\sigma, \rho\} \subseteq \mathbf{\Sigma}_3$ over the action model $\mathbf{\Sigma_3}$ from Example 5. The program Γ represents *the action of "Bob taking a peek at the coin", without any specification of which face he is seeing*. Although expressed in a non-deterministic manner (as a collection of two possible actions, σ and ρ), this program corresponds is in fact *deterministic*, since in each possible state only one of the actions σ or ρ can happen: there is no state satisfying both **H** and **T**. The whole set Σ gives another doxastic program, one that is really non-deterministic: it represents the non-deterministic choice of Bob between taking a peek and not taking it.

Appearance of actions and their revision: Examples. As an example of an agent's "theory" about an action, consider the appearance of action ρ to Alice: $\rho_a = \{\tau\}$. Indeed, if ρ happens (Bob taking a peek and sees the coin is Tails up), Alice believes that τ (i.e., nothing) is happening: this is the "apparent action", as far as Alice is concerned. As an example of a "revised theory" about an action, consider the conditional appearance ρ_a^Γ of ρ to Alice given the program $\Gamma = \{\sigma, \rho\}$ introduced above. It is easy to see that we have $\rho_a^\Gamma = \{\sigma\}$. This captures our intuitions about Alice's revised theory: if, while ρ was happening, she were told that Bob took a peek (i.e., she'd revise with Γ), then she would believe that he saw the coin lying Heads up (i.e., that σ happened).

Example 6 (Successful Lying). Suppose now that, *after* the previous action, i.e., after we arrived in the situation described in Example 3, Bob sneakily announces: "I took a peek and saw the coin was lying *Tails up*". We formalize the content of this announcement as $K_b\mathbf{T}$, i.e., saying that "Bob knows the coin is lying Tails up". This is a *public announcement*, but *not a truthful one* (though it does convey some truthful information): it is a *lie*! We assume it is in fact a *successful lie*: it is common

knowledge that, even after Bob admitted having taken a peek, Alice still believes him. This action is given by the *left node* in the following model $\mathbf{\Sigma}_4$:



The Action-Priority Update

We are ready to define our *update operation*, representing the way an action from a (action) plausibility model $\mathbf{\Sigma} = (\Sigma, \leq_a, \text{pre})_{a\in\mathcal{A}}$ "acts" on an input-state from a given (state) plausibility model $\mathbf{S} = (S, \leq_a, \|\cdot\|)_{a\in\mathcal{A}}$. We denote the updated state model by $\mathbf{S}\otimes\mathbf{\Sigma}$, and call it the *update product* of the two models. The construction is similar to a point to the one in Baltag et al. (1998) and Baltag and Moss (2004), and thus also somewhat similar to the ones in Aucher (2003) and van Ditmarsch (2005). In fact, the set of updated states, the updated valuation and the updated indistinguishability relation are *the same* in these constructions. The main difference lies in our definition of the *updated plausibility relation*, via the *Action Priority Rule*.

Updating Single-Agent Models: The Anti-lexicographic Order

To warm up, let us first define the update product for the single-agent case. Let $\mathbf{S} = (S, \leq, \|\cdot\|)$ be a single-agent plausibility state model and let $\mathbf{\Sigma} = (\Sigma, \leq, \text{pre})$ be a single-agent plausibility action model.

We represent the *states of the updated model* $\mathbf{S}\otimes\mathbf{\Sigma}$ as pairs (s, σ) of input-states and actions, i.e., as elements of the Cartesian product $S \times \Sigma$. This reflects that the basic actions in our action models are assumed to be *deterministic*: For a given input-state and a given action, there can only be at most one output-state. More specifically, we select the pairs which are *consistent, in the sense that the input-state satisfies the precondition of the action*. This is natural: the precondition of an action is a specification of its domain of applicability. So the *set of states* of $\mathbf{S}\otimes\mathbf{\Sigma}$ is taken to be

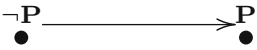
$$S \otimes \Sigma := \{(s, \sigma) : s \models_{\mathbf{S}} \text{pre}(\sigma)\}.$$

The *updated valuation* is essentially given by the *original valuation* from the input-state model: For all $(s, \sigma) \in S \otimes \Sigma$, we put $(s, \sigma) \models p$ iff $s \models p$. This "conservative" way to update the valuation expresses the fact that we only consider here actions that are *"purely doxastic"*, i.e., pure "belief changes", that do not affect the ontic "facts" of the world (captured here by atomic sentences).

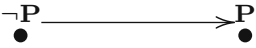
We still need to define the updated plausibility relation. To motivate our definition, we first consider two examples:

Example 7 (A Sample Case). Suppose that we have two states $s, s' \in \mathbf{S}$ such that $s < s'$, $s \models \neg\mathbf{P}$, $s' \models \mathbf{P}$. This means that, if given the supplementary information that the real state is either s or s', the agent believes $\neg\mathbf{P}$:

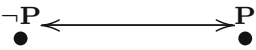
Suppose then an event happens, in whose model there are two actions σ, σ' such that $\sigma > \sigma'$, $\mathrm{pre}_\sigma = \neg\mathbf{P}$, $\mathrm{pre}_{\sigma'} = \mathbf{P}$. In other words, if given the information that either σ or σ' is happening, the agent believes that σ' is happening, i.e., she believes that $\mathbf{P}$ is learnt. This part of the model behaves just like a *soft public announcement* of $\mathbf{P}$:



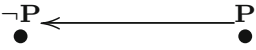
Naturally, we expect the agent to *change her belief* accordingly, i.e., her updated plausibility relation on states should now go the other way:



Example 8 (A Second Sample Case). Suppose the initial situation was the same as above, but now the two actions σ, σ' are assumed to be equi-plausible: $\sigma \cong \sigma'$. This is a *completely unreliable announcement of* $\mathbf{P}$, in which the veracity and the falsity of the announcement are equally plausible alternatives:



In the AGM paradigm, it is natural to expect the agents to *keep their original beliefs* unchanged after this event:



The anti-lexicographic order. Putting the above two sample cases together, we conclude that the updated plausibility relation should be the *anti-lexicographic preorder relation* induced on pairs $(s, \sigma) \in S \times \Sigma$ by the preorders on $\mathbf{S}$ and on $\mathbf{\Sigma}$, i.e.:

$$(s, \sigma) \leq (s', \sigma') \quad \text{iff:} \quad \text{either } \sigma < \sigma', \text{ or else } \sigma \cong \sigma' \text{ and } s \leq s'.$$

In other words, the updated plausibility order gives *"priority" to the action plausibility relation, and apart from this it keeps as much as possible the old order*. This reflects our commitment to an AGM-type of revision, in which the new information has priority over old beliefs. The "actions" represent here the "new information", although (unlike in AGM) this information comes in *dynamic form* (as action plausibility order), and so it is not fully reducible to its propositional content (the action's precondition). In fact, this is a generalization of one of the belief-revision policies encountered in the literature (the so-called "*maximal-Spohn revision*"). But, in the context of our qualitative (conditional) interpretation of plausibility models, we will argue below that this is essentially the only reasonable option.

Updating Multi-agent Models: The General Case

In the multi-agent case, the construction of *the updated state space and updated valuation is the same as above*. But for the updated plausibility relation we need to take into account *a third possibility*: the case when either the initial states or the actions are *distinguishable*, belonging to *different information cells*.

Example 9 (A Third Sample Case). Suppose that we have two states $s, s' \in \mathbf{S}$ such that $s \models \neg\mathbf{P}$, $s' \models \mathbf{P}$, but $s \not\sim_a s'$ are *distinguishable* (i.e., non-comparable):

This means that, if given the supplementary information that the real state is either s or s', the agent immediately *knows* which of the two is the real states, and thus *she knows whether* $\mathbf{P}$ *holds or not*. It is obvious that, after any of the actions considered in the previous two examples, a perfect-recall agent *will continue to know* whether $\mathbf{P}$ held or not, and so *the output-states after* σ *and* σ' *will still be distinguishable (non-comparable)*.

The "Action-Priority" Rule. Putting this together with the other sample cases, we obtain our update rule, in full generality:

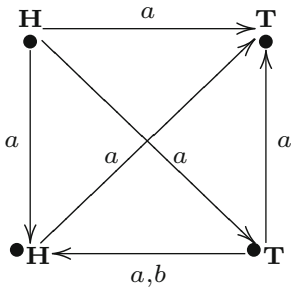
$$(s, \sigma) \leq_a (s', \sigma') \text{ iff either } \sigma <_a \sigma' \text{ and } s \sim_a s', \text{ or else } \sigma \cong_a \sigma' \text{ and } s \leq_a s'$$

We regard this construction as the most natural analogue in a belief-revision context of the similar notion in Baltag and Moss (2004) and Baltag et al. (1998). Following a suggestion of Johan van Benthem, we call this the Action-Priority Update Rule.

Sanity check: Examples 2 and 3 revisited. To check the correctness of our update operation, take first the update product $\mathbf{S} \otimes \mathbf{\Sigma}_2$ of the model $\mathbf{S}$ in Example 1 from the previous section with the action model $\mathbf{\Sigma}_2$ in Example 4 from the previous section. As predicted, the resulting state model is isomorphic to the model $\mathbf{W}$ from

Example 2. Similarly, if $\mathbf{\Sigma}_3$ is the action model from Example 5, then we can see that the product $\mathbf{S}\otimes\mathbf{\Sigma}_3$ is isomorphic to the state model $\mathbf{S}'$ from Example 3.

"In-sanity check": Successful lying. Applying the action model $\mathbf{\Sigma}_4$ in Example 6, representing the "successful lying" action, to the state model $\mathbf{S}'$ from Example 3, we obtain indeed the intuitively correct output of "successful lying", namely the following model $\mathbf{S}'\otimes\mathbf{\Sigma}_4$:



Interpretation. As its name makes explicit, the Action-Priority Rule gives "priority" to the *action* plausibility relation. This is not an arbitrary choice, but it is motivated by our specific interpretation of action models, as embodied in our Motto above: *beliefs about changes* (i.e., the action plausibility relations) *are nothing but ways to encode changes of belief* (i.e., reversals of the original plausibility order). So *the (strict) order on actions encodes changes of order on states*. The Action-Priority Rule is a consequence of this interpretation: it just says that a strong plausibility order $\sigma <_a \sigma'$ on actions corresponds indeed to a change of ordering, (from whatever the ordering was) between the original (indistinguishable) input-states $s \sim_a s'$, to the order $(s,\sigma) <_a (s',\sigma')$ between output-states; while equally plausible actions $\sigma \cong_a \sigma'$ will leave the initial ordering unchanged: $(s,\sigma) \leq_a (s',\sigma')$ iff $s \leq_a s'$. Giving priority to action plausibility does not in any way mean that the agent's belief in actions is stronger than her belief in states; it just captures the fact that, at the time of updating with a given action, *the belief about the action is what is actual, it is the current belief about what is going on, while the beliefs about the input-states are in the past*.[19]

In a nutshell: *the doxastic action is the one that changes the initial doxastic state, and not vice-versa*. The belief update induced by a given action is nothing but an update with the (presently) believed action. If the believed action σ requires the agent to revise some past beliefs, then so be it: this is the whole point of believing σ, namely to use it to revise one's past beliefs. For example, in a successful lying, the action plausibility relation makes the hearer believe that the speaker is telling

[19]Of course, *at a later moment*, the above-mentioned belief about action (*now* belonging to the past) might be itself revised. But this is another, *future update*.

the truth; so she'll accept this message (unless contradicted by her knowledge), and change her past beliefs appropriately: this is what makes the lying successful.

Action-priority update generalizes product update. Recall the definition of the epistemic indistinguishability relation $\sim_a$ in a plausibility model: $s \sim_a s'$ iff either $s \leq_a s'$ or $s' \leq_a s$. It is easy to see that the Action Priority Update implies the familiar update rule from Baltag et al. (1998) and Baltag and Moss (2004), known in Dynamic Epistemic Logic as the "product update":

$$(s, \sigma) \sim_a (s', \sigma') \text{ iff } s \sim_a s' \text{ and } \sigma \sim_a \sigma'.$$

Program transitions. For every state model $\mathbf{S}$, every program $\Gamma \subseteq \Sigma$ over an action model $\boldsymbol{\Sigma}$ induces a transition relation $\overset{\Gamma}{\rightarrow}_{\mathbf{S}} \subseteq \mathbf{S} \times (S \otimes \Sigma)$ from $\mathbf{S}$ to $\mathbf{S} \otimes \boldsymbol{\Sigma}$, given by:

$$s \overset{\Gamma}{\rightarrow}_{\mathbf{S}} (s', \gamma) \text{ iff } s = s', (s, \gamma) \in S \otimes \Sigma \text{ and } \gamma \in \Gamma .$$

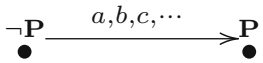
Simulating Various Belief-Revision Policies

We give here three examples of *multi-agent belief-revision policies* that can be simulated by our product update: *truthful public announcements of "hard facts"*, *"lexicographic update"* and *"conservative upgrade"*. They were all introduced by van Benthem in (2007), as multi-agent versions of revision operators previously considered by Rott (1989) and others.

Public announcements of "hard facts". A *truthful public announcement* $!\mathbf{P}$ *of some "hard fact"* $\mathbf{P}$ is not really about belief revision, but about the learning of *certified true information*: it establishes *common knowledge* that $\mathbf{P}$ was the case. This is the action described in van Benthem (2007) as (public) "belief change under hard facts". As an operation on models, this is described in van Benthem (2007) as taking any state model $\mathbf{S}$ and *deleting all the non-*$\mathbf{P}$ *states, while keeping the same indistinguishability and plausibility relations between the surviving states*. In our setting, the corresponding action model consists of only one node, labeled with $\mathbf{P}$. It is easy to see that the above operation on models can be exactly "simulated" by taking the anti-lexicographic product update with this one-node action model.

Public announcements of "soft facts": The "lexicographic upgrade". To allow for "soft" belief revision, an operation $\Uparrow\mathbf{P}$ was introduced in van Benthem (2007), essentially adapting to public announcements the 'lexicographic' policy for belief revision described in Rott (1989). This operation, called "lexicographic update" consists of changing the current plausibility order on any given state model as follows: *every* $\mathbf{P}$*-world becomes "better" (more plausible) than all* $\neg\mathbf{P}$*-worlds in*

*the same information cell, and within the two zones (***P** *and* $\neg$**P***), the old ordering remains*. In our setting, this action corresponds to the following local plausibility action model:



Taking the anti-lexicographic update product with this action will give an exact "simulation" of the lexicographic upgrade operation.

"Conservative upgrade". The operation $\uparrow$**P** of "conservative upgrade", also defined in van Benthem (2007), changes any model as follows: *in every information cell, the best* **P***-worlds become better than all the worlds in that cell* (i.e., in every cell the most plausible **P**-states become the most plausible overall in that cell), *and apart from that, the old order remains*. In the case of a system *with only one agent*, it is easy to see that we have $\uparrow \mathbf{P} = \Uparrow(*_a \mathbf{P})$, where $*_a$ is the unary "revision modality" introduced in the previous section. In the case of a set $\mathcal{A} = \{1, \cdots, n\}$ with $n > 1$ agents, we can simulate $\uparrow$**P** using a model with 2^n actions $\{\uparrow_I \mathbf{P}\}_{I \subseteq \mathcal{A}}$, with

$$\text{pre}_{\uparrow_I \mathbf{P}} = \bigwedge_{i \in I} *_i \mathbf{P} \wedge \bigwedge_{j \notin I} \neg *_j \mathbf{P},$$

$$\uparrow_I \mathbf{P} \leq_k \uparrow_J \mathbf{P} \quad \text{iff} \quad J \cap \{k\} \subseteq I.$$

Operations on Doxastic Programs

First, we introduce *dynamic modalities*, capturing *the "weakest precondition" of a program* Γ. These are the natural analogues of the PDL modalities for our program transition relations $\overset{\Gamma}{\rightarrow}$ between models.

Dynamic modalities. Let $\mathbf{\Sigma}$ be some action plausibility model and $\Gamma \subseteq \Sigma$ be a doxastic model over $\mathbf{\Sigma}$. For every doxastic proposition **P**, we define a doxastic proposition $[\Gamma]\mathbf{P}$ given by

$$([\Gamma]\mathbf{P})_{\mathbf{S}} := [\overset{\Gamma}{\rightarrow}_{\mathbf{S}}]\mathbf{P}_{\mathbf{S}} = \{s \in S : \forall t \in S \otimes \Sigma \, (s \overset{\Gamma}{\rightarrow}_{\mathbf{S}} t \Rightarrow t \models_{\mathbf{S} \otimes \mathbf{\Sigma}} \mathbf{P})\}.$$

For *basic doxastic actions* $\sigma \in \Sigma$, we define the dynamic modality $[\sigma]$ via the above-mentioned identification of actions σ with singleton programs $\{\sigma\}$:

$$([\sigma]\mathbf{P})_{\mathbf{S}} := ([\{\sigma\}]\mathbf{P})_{\mathbf{S}} = \{s \in S : \text{ if } (s, \sigma) \in \mathbf{S} \otimes \mathbf{\Sigma} \text{ then } (s, \sigma) \in \mathbf{P}_{\mathbf{S} \otimes \mathbf{\Sigma}}\}.$$

The dual (Diamond) modalities are defined as usually: $\langle \Gamma \rangle \mathbf{P} := \neg[\Gamma]\neg\mathbf{P}$.

We can now introduce operators on doxastic programs that are the analogues of the *regular operations* of PDL.

Sequential composition. The *sequential composition* $\mathbf{\Sigma};\mathbf{\Delta}$ of two action plausibility models $\mathbf{\Sigma} = (\Sigma, \leq_a, \text{pre})$, $\mathbf{\Delta} = (\Delta, \leq_a, \text{pre})$ is defined as follows:

- the set of basic actions is the Cartesian product $\Sigma \times \Delta$
- the preconditions are given by $\text{pre}_{(\sigma,\delta)} := \langle\sigma\rangle\, \text{pre}_\delta$
- the plausibility order is given by putting $(\sigma, \delta) \leq_a (\sigma', \delta')$ iff: either $\sigma <_a \sigma'$ and $\delta \sim_a \delta'$, or else $\sigma \cong_a \sigma'$ and $\delta \leq_a \delta'$.

We think of (σ, δ) as the action of *performing first* σ *then* δ, and thus use the notation

$$\sigma;\delta := (\sigma, \delta).$$

We can extend this notation to doxastic programs, by defining the *sequential composition of programs* $\Gamma \subseteq \Sigma$ and $\Lambda \subseteq \Delta$ to be a program $\Gamma;\Lambda \subseteq \Sigma;\Delta$ over the action model $\mathbf{\Sigma};\mathbf{\Delta}$, given by:

$$\Gamma;\Lambda := \{(\gamma, \lambda) : \gamma \in \Gamma, \lambda \in \Lambda\}.$$

It is easy to see that this behaves indeed like a sequential composition:

Proposition 12. *For every state plausibility model* $\mathbf{S}$*, action plausibility models* $\mathbf{\Sigma}$ *and* $\mathbf{\Delta}$*, and programs* $\Gamma \subseteq \Sigma$*,* $\Lambda \subseteq \Delta$*, we have the following:*

1. *The state plausibility models* $(\mathbf{S} \otimes \mathbf{\Sigma}) \otimes \mathbf{\Delta}$ *and* $\mathbf{S} \otimes (\mathbf{\Sigma};\mathbf{\Delta})$ *are isomorphic, via the canonical map* $F : (S \otimes \Sigma) \otimes \Delta \to S \otimes (\Sigma;\Delta)$ *given by*

$$F(\,((s, \sigma), \delta)\,) \;:=\; (s, (\sigma, \delta)).$$

2. *The transition relation for the program* $\Gamma;\Delta$ *is the relational composition of the transition relations for* Γ *and for* Δ *and of the isomorphism map F:*

$s \overset{\Gamma;\Delta}{\to}_{\mathbf{S}} s'$ iff there exist $w, t \in S \otimes \Sigma$ such that

$$s \overset{\Gamma}{\to}_{\mathbf{S}} w \overset{\Delta}{\to}_{\mathbf{S}\otimes\mathbf{\Sigma}} t \text{ and } F(t) = s'.$$

Union (non-deterministic choice). If $\mathbf{\Sigma} = (\Sigma, \leq_a, \text{pre})$ and $\mathbf{\Delta} = (\Delta, \leq'_a, \text{pre}')$ are two action plausibility models, their *disjoint union* $\mathbf{\Sigma} \sqcup \mathbf{\Delta}$ is simply given by taking as set of states the disjoint union $\Sigma \sqcup \Delta$ of the two sets of states, taking as plausibility order the disjoint union $\leq_a \sqcup \leq'_a$ and as precondition map the disjoint union $\text{pre} \sqcup \text{pre}'$ of the two precondition maps. If $\Gamma \subseteq \Sigma$ and $\Lambda \subseteq \Delta$ are doxastic programs over the two models, we define their *union* to be the program over the model $\mathbf{\Sigma} \sqcup \mathbf{\Delta}$ given by the disjoint union $\Gamma \sqcup \Lambda$ of the sets of actions of the two programs.

Again, it is easy to see that *this behaves indeed like a non-deterministic choice operator*:

Proposition 13. *Let* $i_1 : \Sigma \to \Sigma \sqcup \Delta$ *and* $i_2 : \Delta \to \Sigma \sqcup \Delta$ *be the two canonical injections. Then the following are equivalent:*

- $s \overset{\Gamma \sqcup \Delta}{\to}_{\mathbf{S}} s'$
- *there exists* t *such that:*

$$\text{either } s \overset{\Gamma}{\to}_{\mathbf{S}} t \text{ and } i_1(t) = s', \text{ or else } s \overset{\Delta}{\to}_{\mathbf{S}} t \text{ and } i_2(t) = s'.$$

Other operators. *Arbitrary unions* $\bigsqcup_i \Gamma_i$ can be similarly defined, and then one can define *iteration* $\Gamma^* := \bigsqcup_i \Gamma^i$ (where $\Gamma^0 = !\top$ and $\Gamma^{i+1} = \Gamma; \Gamma^i$).

The Laws of Dynamic Belief Revision

The "laws of dynamic belief revision" are the fundamental equations of Belief Dynamics, allowing us *to compute future doxastic attitudes from past ones*, given the doxastic events that happen in the meantime. In modal terms, these can be stated as "reduction laws" for inductively computing dynamic modalities $[\Gamma]\mathbf{P}$, by reducing them to modalities $[\Gamma']\mathbf{P}'$ in which either the propositions $\mathbf{P}'$ or the programs Γ' have *lower complexity*.

The following immediate consequence of the definition of $[\Gamma]\mathbf{P}$ allows us to reduce modalities for non-deterministic programs Γ to the ones for their deterministic resolutions $\gamma \in \Gamma$:

Deterministic Resolution Law. For every program $\Gamma \subseteq \Sigma$, we have

$$[\Gamma]\mathbf{P} \;=\; \bigwedge_{\gamma \in \Gamma} [\gamma]\mathbf{P}\,.$$

So, for our other laws, we can restrict ourselves to *basic actions* in Σ.

The Action-Knowledge Law. For every action $\sigma \in \Sigma$, we have:

$$[\sigma]K_a\mathbf{P} \;=\; \mathrm{pre}_\sigma \to \bigwedge_{\sigma' \sim_a \sigma} K_a[\sigma']\mathbf{P}\,.$$

This Action-Knowledge Law is essentially the same as in Baltag et al. (1998) and Baltag and Moss (2004): *a proposition* $\mathbf{P}$ *will be known after a doxastic event iff, whenever the event can take place, it is known that* $\mathbf{P}$ *will become true after all events that are indistinguishable from the given one*.

The Action-Safe-Belief Law. For every action $\sigma \in \Sigma$, we have:

$$[\sigma]\Box_a \mathbf{P} \;=\; \mathrm{pre}_\sigma \to \bigwedge_{\sigma' <_a \alpha} K_a[\sigma']\mathbf{P} \wedge \bigwedge_{\sigma'' \cong_a \sigma} \Box_a[\sigma'']\mathbf{P}\,.$$

This law embodies the essence of the Action-Priority Rule: *a proposition* **P** *will be safely believed after a doxastic event iff, whenever the event can take place, it is known that* **P** *will become true after all more plausible events and in the same time it is safely believed that* **P** *will become true after all equi-plausible events*.

Since we took knowledge and safe belief as the basis of our static logic $K\Box$, the above two laws are the "fundamental equations" of our theory of dynamic belief revision. But note that, as a consequence, one can obtain *derived laws for (conditional) belief* as well. Indeed, using the above-mentioned characterization of conditional belief in terms of K and $\Box$, we obtain the following:

The Derived Law of Action-Conditional-Belief. For every action $\sigma \in \Sigma$, we have:

$$[\sigma]B_a^{\mathbf{P}}\mathbf{Q} \;=\; \mathrm{pre}_\sigma \to \bigvee_{\Gamma \subseteq \Sigma} \Big(\bigwedge_{\gamma \in \Gamma} \tilde{K}_a\langle\gamma\rangle\mathbf{P} \wedge \bigwedge_{\gamma' \notin \Gamma} \neg\tilde{K}_a\langle\gamma'\rangle\mathbf{P} \wedge B_a^{\langle\sigma_a^\Gamma\rangle\mathbf{P}}[\sigma_a^\Gamma]\mathbf{Q} \Big)\,.$$

This derived law, a version of which was first introduced in Baltag and Smets (2006c) (where it was considered a fundamental law), allows us to predict future conditional beliefs from current conditional beliefs.

To explain the meaning of this law, we re-state it as follows: For every $s \in \mathbf{S}$ and $\sigma \in \mathbf{\Sigma}$, we have:

$$s \models [\sigma]B_a^{\mathbf{P}}\mathbf{Q} \;\text{ iff }\; s \models \mathrm{pre}_\sigma \to B_a^{\langle\sigma_a^\Gamma\rangle\mathbf{P}}[\sigma_a^\Gamma]\mathbf{Q}\,,$$
$$\text{where }\; \Gamma = \{\gamma \in \Sigma : s \models_{\mathbf{S}} \tilde{K}_a\langle\gamma\rangle\mathbf{P}\}\,.$$

It is easy to see that this "local" (state-dependent) version of the reduction law is equivalent to the previous (state-independent) one. The set Γ encodes the extra information about the current action that is given to the agent by the context s and by the post-condition **P**; while σ_a^Γ is the action's *post-conditional contextual appearance*, i.e., the way it appears to the agent in the view of this extra-information Γ. Indeed, a given action might "appear" differently in a given context (i.e., at a state s) than it does in general: the information possessed by the agent at the state s might imply the negation of certain actions, hence their impossibility; this information will then be used to revise the agent's beliefs about the actions, obtaining her contextual beliefs. Moreover, in the presence of further information (a "post-condition" **P**), this appearance might again be revised. The "post-conditional contextual appearance" is the result of this double revision: the agent's belief about action σ is revised with the information given to her by the context s and the post-condition **P**. This information

is encoded in a set $\Gamma = \{\gamma \in \Sigma : s \models_{\mathbf{S}} \tilde{K}_a\langle\gamma\rangle\mathbf{P}\}$ of "admissible" actions: the actions for which the agent considers epistemically possible (at s) that they can be performed and they can achieve the post-condition $\mathbf{P}$. The "post-conditional contextual appearance" σ_a^Γ of action σ captures the agent's revised theory about σ after revision with the relevant information Γ.

So the above law says that: *the agent's future conditional beliefs* $[\sigma]B_a^{\mathbf{P}}$ *can be predicted, given that action* σ *happens, by her current conditional beliefs* $B_a^{\langle\sigma_a^\Gamma\rangle\mathbf{P}}[\sigma_a^\Gamma]$ *about what will be true after the apparent action* σ_a^Γ (as it appears in the given context and in the view of the given post-condition $\mathbf{P}$), *beliefs conditioned on the information* $(\langle\sigma_a^\Gamma\rangle\mathbf{P})$ *that the apparent action* σ_a^Γ *actually can lead to the fulfillment of the post-condition* $\mathbf{P}$.

Special cases. As special cases of the Action-Conditional-Belief Law, we can derive *all the reduction laws* in van Benthem (2007) for (conditional) belief after the events $!\mathbf{P}$, $\Uparrow\mathbf{P}$ and $\uparrow\mathbf{P}$:

$$[!\mathbf{P}]B_a^{\mathbf{Q}}\mathbf{R} = \mathbf{P} \to B_a^{\mathbf{P}\wedge[!\mathbf{P}]\mathbf{Q}}[!\mathbf{P}]\mathbf{R}\,,$$

$$[\Uparrow\mathbf{P}]B_a^{\mathbf{Q}}\mathbf{R} = \big(\tilde{K}_a^{\mathbf{P}}[\Uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{\mathbf{P}\wedge[\Uparrow\mathbf{P}]\mathbf{Q}}[\Uparrow\mathbf{P}]\mathbf{R}\big) \vee \big(\neg\tilde{K}_a^{\mathbf{P}}[\Uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{[\Uparrow\mathbf{P}]\mathbf{Q}}[\Uparrow\mathbf{P}]\mathbf{R}\big)\,,$$

$$[\uparrow\mathbf{P}]B_a^{\mathbf{Q}}\mathbf{R} = \big(\tilde{B}_a^{\mathbf{P}}[\uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{\mathbf{P}\wedge[\uparrow\mathbf{P}]\mathbf{Q}}[\uparrow\mathbf{P}]\mathbf{R}\big) \vee \big(\neg\tilde{B}_a^{\mathbf{P}}[\uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{[\uparrow\mathbf{P}]\mathbf{Q}}[\uparrow\mathbf{P}]\mathbf{R}\big)\,,$$

where

$$K_a^{\mathbf{P}}\mathbf{Q} := K_a(\mathbf{P}\to\mathbf{Q})\,, \qquad \tilde{K}_a^{\mathbf{P}}\mathbf{Q} := \neg K_a^{\mathbf{P}}\neg\mathbf{Q}\,, \qquad \tilde{B}_a^{\mathbf{P}}\mathbf{Q} := \neg B_a^{\mathbf{P}}\neg\mathbf{Q}\,.$$

Laws for other doxastic attitudes. The *equi-plausibility modality behaves dynamically "almost" like knowledge*, while *the strict plausibility modality behaves like safe belief*, as witnessed by the following laws:

$$[\sigma][\cong_a]\mathbf{P} = \mathrm{pre}_\sigma \to \bigwedge_{\sigma'\cong_a\sigma}[\cong_a][\sigma']\mathbf{P}\,,$$

$$[\sigma][>_a]\mathbf{P} = \mathrm{pre}_\sigma \to \bigwedge_{\sigma'<_a\alpha} K_a[\sigma']\mathbf{P} \wedge \bigwedge_{\sigma''\cong_a\sigma}[>_a][\sigma'']\mathbf{P}\,.$$

From these, we can *derive laws for all the other doxastic attitudes* above.

The Logic of Doxastic Actions

The problem of finding a general syntax for action models has been tackled in various ways by different authors. Here we use the *action-signature approach* from Baltag and Moss (2004).

Signature. A doxastic *action signature* is a *finite plausibility frame* $\mathbf{\Sigma}$, together with an *ordered list without repetitions* $(\sigma_1, \ldots, \sigma_n)$ of some of the elements of $\mathbf{\Sigma}$. The elements of $\mathbf{\Sigma}$ are called *action types*. A type σ is called *trivial* if it is *not* in the above list.

Example 10. The *"hard" public announcement signature* HardPub is a singleton frame, consisting of one action type !, identity as the order relation, and the list (!).

The *"soft" public announcement signature* SoftPub is a two-point frame, consisting of types $\Uparrow$ and $\Downarrow$, with $\Downarrow <_a \Uparrow$ for all agents a, and the list $(\Uparrow, \Downarrow)$.

Similarly, one can define the signatures of *fully private announcements with n alternatives*, *private "fair-game" announcements, conservative upgrades* etc. As we will see below, *there is no signature of "successful (public) lying": public lying actions fall under the type of "soft" public announcements*, so they are generated by that signature.

Languages. For each action signature $(\mathbf{\Sigma}, (\sigma_1, \ldots, \sigma_n))$, the language $L(\mathbf{\Sigma})$ consists of a set of *sentences* φ and a set of *program terms* π, defined by simultaneous recursion:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid \Box_a\varphi \mid [\pi]\varphi$$
$$\pi ::= \sigma\varphi_1 \ldots \varphi_n \mid \pi \sqcup \pi \mid \pi;\pi$$

where $p \in \Phi$, $a \in \mathcal{A}$, $\sigma \in \Sigma$, and $\sigma\varphi_1 \ldots \varphi_n$ is an expression consisting of σ and a string of n sentences, where n is the length of the list $(\sigma_1, \ldots, \sigma_n)$.

Syntactic action model. The expressions of the form $\sigma\vec{\varphi}$ are called *basic programs*. The preorders on $\mathbf{\Sigma}$ induce in a natural way preorders on the basic programs in $L(\mathbf{\Sigma})$:

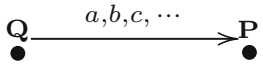
$$\sigma\vec{\varphi} \leq_a \sigma\vec{\psi} \text{ iff } \sigma \leq_a \sigma' \text{ and } \vec{\varphi} = \vec{\psi}.$$

The given listing can be used to assign syntactic preconditions for basic programs, by putting: $\mathrm{pre}_{\sigma_i\vec{\varphi}} := \varphi_i$, and $\mathrm{pre}_{\sigma\vec{\varphi}} := \top$ (the trivially true sentence) if σ is not in the listing. Thus, the basic programs of the form $\sigma\vec{\varphi}$ form a *"syntactic plausibility model"* $\mathbf{\Sigma}\vec{\varphi}$; i.e., every given interpretation $\|\cdot\| : L(\mathbf{\Sigma}) \to$ Prop of sentences as doxastic propositions will convert this syntactic model into a "real" (semantic) plausibility model, called $\mathbf{\Sigma}\|\vec{\varphi}\|$.

Action models induced by a signature. For a given signature $\mathbf{\Sigma}$, let $(\sigma_1, \ldots, \sigma_n)$ be its list of non-trivial types, and let $\vec{\mathbf{P}} = (\mathbf{P}_1, \ldots, \mathbf{P}_n)$ be a matching list of doxastic propositions. The *action model generated by the signature* $\mathbf{\Sigma}$ *and the list of propositions* $\vec{\mathbf{P}}$ is the model $\mathbf{\Sigma}\vec{\mathbf{P}}$, having $\mathbf{\Sigma}$ as its underlying action frame and having a precondition map given by: $\mathrm{pre}_{\sigma_i} = \mathbf{P}_i$, for non-trivial types σ_i; and $\mathrm{pre}_{\sigma} = \top$ (the trivially true proposition), for trivial types σ. When referring to σ as an *action* in $\mathbf{\Sigma}\vec{\mathbf{P}}$, we will denote it by $\sigma\vec{\mathbf{P}}$, to distinguish it from the action *type* $\sigma \in \Sigma$.

We can obviously extend this construction to *sets of action types*: given a signature $\mathbf{\Sigma}$ and a list $\vec{\mathbf{P}} = (\mathbf{P}_1, \ldots, \mathbf{P}_n)$, every set $\Gamma \subseteq \Sigma$ gives rise to a doxastic program $\Gamma\vec{\mathbf{P}} := \{\sigma\vec{\mathbf{P}} : \sigma \in \Sigma\} \subseteq \Sigma\vec{\mathbf{P}}$.

Example 11. The action model of a hard public announcement $!\,\mathbf{P}$ is generated as $!(\mathbf{P})$ by the hard public announcement signature $\mathsf{HardPub} = \{!\}$ and the list $(\mathbf{P})$. Similarly, the action model $\mathsf{SoftPub}(\mathbf{P})$ generated by the *soft* public announcement signature $\mathsf{SoftPub}$ and a list $(\mathbf{P}, \mathbf{Q})$ of two propositions consists of two actions $\Uparrow(\mathbf{P}, \mathbf{Q})$ and $\Downarrow(\mathbf{P}, \mathbf{Q})$, with $\Uparrow(\mathbf{P}, \mathbf{Q}) <_a \Downarrow(\mathbf{P}, \mathbf{Q})$, $\mathrm{pre}_{\Uparrow(\mathbf{P},\mathbf{Q})} = \mathbf{P}$ and $\mathrm{pre}_{\Downarrow(\mathbf{P},\mathbf{Q})} = \mathbf{Q}$:



This represents an event during which all agents share a common belief that $\mathbf{P}$ is announced; but they might be wrong and maybe $\mathbf{Q}$ was announced instead. However, it is common knowledge that either $\mathbf{P}$ or $\mathbf{Q}$ was announced.

Successful (public) lying $\mathrm{Lie}\,\mathbf{P}$ (*by an anonymous agent, falsely announcing* $\mathbf{P}$) can now be expressed as $\mathrm{Lie}\,\mathbf{P} := \Downarrow(\mathbf{P}, \neg\mathbf{P})$. The *truthful* soft announcement is $\mathrm{True}\,\mathbf{P} := \Uparrow(\mathbf{P}, \neg\mathbf{P})$. Finally, the soft public announcement (lexicographic update) $\Uparrow\mathbf{P}$, as previously defined, is given by the non-deterministic union $\Uparrow\mathbf{P} := \mathrm{True}\,\mathbf{P} \sqcup \mathrm{Lie}\,\mathbf{P}$.

Semantics. We define by simultaneous induction two *interpretation maps*, one taking sentences φ into doxastic propositions $\|\varphi\| \in \mathsf{Prop}$, the second taking program terms π into doxastic programs $\|\pi\|$ over some plausibility frames. The inductive definition uses the obvious semantic clauses. For programs: $\|\sigma\vec{\varphi}\|$ is the action $\sigma\|\vec{\varphi}\|$ (or, more exactly, the singleton program $\{\sigma\|\vec{\varphi}\|\}$ over the frame $\mathbf{\Sigma}\|\vec{\varphi}\|$), $\|\pi \sqcup \pi'\| := \|\pi\| \sqcup \|\pi'\|$, $\|\pi; \pi'\| := \|\pi\|; \|\pi'\|$. For sentences: $\|p\|$ is as given by the valuation, $\|\neg\varphi\| := \neg\|\varphi\|$, $\|\varphi \wedge \psi\| := \|\varphi\| \wedge \|\psi\|$, $\|K_a\varphi\| := K_a\|\varphi\|$, $\|\Box_a\varphi\| := \Box_a\|\varphi\|$, $\|[\pi]\varphi\| := [\,\|\pi\|\,]\,\|\varphi\|$.

Proof system. In addition to the axioms and rules of the logic $K\Box$, the logic $L(\mathbf{\Sigma})$ includes the following Reduction Axioms:

$$
\begin{aligned}
[\alpha]p &\leftrightarrow \mathrm{pre}_\alpha \to p\\
[\alpha]\neg\varphi &\leftrightarrow \mathrm{pre}_\alpha \to \neg[\alpha]\varphi\\
[\alpha](\varphi \wedge \psi) &\leftrightarrow \mathrm{pre}_\alpha \to [\alpha]\varphi \wedge [\alpha]\psi\\
[\alpha]K_a\varphi &\leftrightarrow \mathrm{pre}_\alpha \to \bigwedge_{\alpha' \sim_a \alpha} K_a[\alpha']\varphi\\
[\alpha]\Box_a\varphi &\leftrightarrow \mathrm{pre}_\alpha \to \bigwedge_{\alpha' <_a \alpha} K_a[\alpha']\varphi \wedge \bigwedge_{\alpha'' \cong_a \alpha} \Box_a[\alpha'']\varphi\\
[\pi \sqcup \pi']\varphi &\leftrightarrow [\pi]\varphi \wedge [\pi']\varphi\\
[\pi; \pi']\varphi &\leftrightarrow [\pi]\,[\pi']\varphi
\end{aligned}
$$

where p is any atomic sentence, π, π' are program terms, α is a *basic* program term in $L(\mathbf{\Sigma})$, pre is the syntactic precondition map defined above, and $\sim_a, <_a, \cong_a$ are respectively the (syntactic) epistemic indistinguishability, the strict plausibility order and the equi-plausibility relation on basic programs.

Theorem 16. *For every signature* $\mathbf{\Sigma}$*, the above proof system for the dynamic logic* $L(\mathbf{\Sigma})$ *is* complete, decidable *and* has the finite model property. *In fact, this dynamic logic has the same expressive power as the "static" logic* $K\Box$ *of knowledge and safe belief.*

Proof (Sketch). The proof is similar to the ones in Baltag and Moss (2004), Baltag et al. (1998), and van Ditmarsch et al. (2007). We use the reduction laws to inductively simplify any formula until it is reduced to a formula of the $K\Box$-logic, then use the completeness of the $K\Box$ logic. Note that this is *not* an induction on subformulas, but (as in Baltag et al. 1998) on an appropriate notion of "complexity" ordering of formulas. Q.E.D.

Current and Future Work, Some Open Questions

In our papers Baltag and Smets (2007a,b), we present a *probabilistic version* of the theory developed here, based on *discrete (finite) Popper-Renyi conditional probability spaces* (allowing for conditionalization on events of non-zero probability, in order to cope with non-trivial belief revisions). We consider subjective probability to be the proper notion of "degree of belief", and we investigate its relationship with the qualitative concepts developed here. We develop a probabilistic generalization of the Action Priority Rule, and show that the logics presented above are *complete for the (discrete) conditional probabilistic semantics*.

We mention here a number of open questions: (1) Axiomatize the full (static) logic of doxastic attitudes introduced in this paper. It can be easily shown that they can all be reduced to the modalities K_a, $[>_a]$ and $[\cong_a]$. There are a number of obvious axioms for the resulting logic $K[>][\cong]$ (note in particular that $[>]$ satisfies the Gödel-Löb formula!), but the completeness problem is still open. (2) Axiomatize the logic of *common safe belief and common knowledge*, and their *dynamic versions*. More generally, explore the logics obtained by adding *fixed points*, or at least "epistemic regular (PDL-like) operations" as in van Benthem et al. (2006b), on top of our doxastic modalities. (3) Investigate the *expressive limits* of this approach *with respect to belief-revision policies*: what policies can be simulated by our update? (4) Extend the work in Baltag and Smets (2007a,b), by investigating and axiomatizing doxastic logics on *infinite* conditional probability models. (5) Extend the logics with *quantitative (probabilistic)* modal operators $B^P_{a,x}Q$ (or $\Box_{a,x}Q$) expressing that *the degree of conditional belief in* Q *given* P (or the *degree of safety* of the belief in Q) is at least x.

Acknowledgements Sonja Smets' contribution to this research was made possible by the post-doctoral fellowship awarded to her by the Flemish Fund for Scientific Research. We thank Johan van Benthem for his insights and help, and for the illuminating discussions we had with him on the topic of this paper. His pioneering work on dynamic belief revision acted as the "trigger" for our own. We also thank Larry Moss, Hans van Ditmarsch, Jan van Eijck and Hans Rott for their most valuable feedback. Finally, we thank the editors and the anonymous referees of the LOFT7-proceedings for their useful suggestions and comments.

During the republication of this paper in 2015, the research of Sonja Smets was funded by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement no.283963.

References

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic, 50*(2), 510–530.

Aucher, G. (2003). *A combined system for update logic and belief revision*. Master's thesis, University of Amsterdam. ILLC Publications MoL-2003-03.

Aumann, R. J. (1999). Interactive epistemology I: Knowledge. *International Journal of Game Theory, 28*(3), 263–300.

Baltag, A. (2002). A logic for suspicious players: Epistemic actions and belief updates in games. *Bulletin of Economic Research, 54*(1), 1–46.

Baltag, A., & Moss, L. S. (2004). Logics for epistemic programs. *Synthese, 139*(2), 165–224.

Baltag, A., Moss, L. S., & Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions. In I. Gilboa (Ed.), *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, (pp. 43–56).

Baltag, A., & Sadrzadeh, M. (2006). The algebra of multi-agent dynamic belief revision. *Electronic Notes in Theoretical Computer Science, 157*(4), 37–56.

Baltag, A., & Smets, S. (2006). Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electronic Notes in Theoretical Computer Science, 165*, 5–21.

Baltag, A., & Smets, S. (2006b) Dynamic belief revision over multi-agent plausibility models. In Bonanno et al. (2006) (pp. 11–24).

Baltag, A., & Smets, S. (2006c). The logic of conditional doxastic actions: A theory of dynamic multi-agent belief revision. In S. Artemov, & Parikh, R. (Eds.), *Proceedings of ESSLLI Workshop on Rationality and Knowledge*, (pp. 13–30). ESSLLI.

Baltag, A., & Smets, S. (2007a). From conditional probability to the logic of doxastic actions. In D. Samet (Ed.), *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, Brussels (pp. 52–61). UCL Presses Universitaires de Louvain.

Baltag, A., & Smets, S. (2007b). Probabilistic dynamic belief revision. In J. F. A. K. van Benthem, S. Ju, & F. Veltman (Eds.), *A Meeting of the Minds: Proceedings of the Workshop on Logic, Rationality and Interaction*, Beijing, 2007 (Texts in computer science, Vol. 8). London: College Publications.

Battigalli, P., & Siniscalchi, M. (2002). Strong belief and forward induction reasoning. *Journal of Economic Theory, 105*(2), 356–391.

Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal logic* (Cambridge tracts in theoretical computer science, Vol. 53). Cambridge: Cambridge University Press.

Board, O. (2002). Dynamic interactive epistemology. *Games and Economic Behaviour, 49*(1), 49–80.

Bonanno, G. (2005). A simple modal logic for belief revision. *Synthese, 147*(2), 193–228.

Bonanno, G., van der Hoek, W., & Wooldridge, M. (Eds.). (2006). *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory (LOFT7)*, University of Liverpool UK.

Friedmann, N., & Halpern, J. Y. (1994). Conditional logics of belief revision. In *Proceedings of the of 12th National Conference on Artificial Intelligence (AAAI-94)*, Seattle, 31 July–4 Aug 1994 (pp. 915–921). Menlo Park: AAAI.

Gärdenfors, P. *Knowledge in flux: Modelling the dynamics of epistemic states*. Gardenfors. 1988, MIT Press, Cambridge/London.

Gerbrandy, J. (1999). Dynamic epistemic logic. In L. S. Moss, J. Ginzburg, & M. de Rijke (Eds.), *Logic, language and information* (Vol. 2, p. 67–84). Stanford: CSLI Publications/Stanford University.

Gerbrandy, J., & Groeneveld, W. (1997). Reasoning about information change. *Journal of Logic, Language and Information, 6*(2), 147–169.

Gerbrandy, J. D. (1999). *Bisimulations on planet Kripke*. PhD thesis, University of Amsterdam. ILLC Publications, DS-1999-01.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis, 23*(6), 121–123.

Gochet, P., & Gribomont, P. (2006). Epistemic logic. In D. M. Gabbay & J. Woods (Eds.), *Handbook of the history of logic* (Vol. 7, p. 99–195). Oxford: Elsevier.

Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic, 17*(2), 157–170.

Hintikka, J. (1962). *Knowledge and belief*. Ithaca: Cornell University Press.

Katsuno, H., & Mendelzon, A. O. (1992). On the difference between updating a knowledge base and revising it. In P. Gärdenfors (Ed.), *Belief revision* (Cambridge tracts in theoretical computer science, pp. 183–203). Cambridge/New York: Cambridge University Press.

Klein, P. (1971). A proposed definition of propositional knowledge. *Journal of Philosophy, 68*(16), 471–482.

Kooi, B. P. (2003). Probabilistic dynamic epistemic logic. *Journal of Logic, Language and Information, 12*(4), 381–408.

Lehrer, K. (1990). *Theory of knowledge*. London: Routledge.

Lehrer, K., & Paxson, T. Jr. (1969). Knowledge: Undefeated justified true belief. *Journal of Philosophy, 66*(8), 225–237.

Meyer, J.-J. Ch. & van der Hoek, W. (1995). *Epistemic logic for AI and computer science* (Cambridge tracts in theoretical computer science, Vol. 41). Cambridge: Cambridge University Press.

Pappas, G., & Swain, M. (Eds.). (1978). *Essays on knowledge and justification*. Ithaca: Cornell University Press.

Plaza, J. A. (1989). Logics of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, & Z. W. Ras (Eds.), *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems Poster Session Program* (pp. 201–216). Oak Ridge National Laboratory, ORNL/DSRD-24.

Rott, H. (1989). Conditionals and theory change: Revisions, expansions, and additions. *Synthese, 81*(1), 91–113.

Rott, H. (2004). Stability, strength and sensitivity: Converting belief into knowledge. *Erkenntnis, 61*(2–3), 469–493.

Ryan, M., & Schobbens, P.-Y. (1997). Counterfactuals and updates as inverse modalities. *Journal of Logic, Language and Information, 6*(2), 123–146.

Segerberg, K. (1998). Irrevocable belief revision in dynamic doxastic logic. *Notre Dame Journal of Formal Logic, 39*(3), 287–306.

Spohn, W. (1988). Ordinal conditional functions: A dynamic theory of epistemic states. In W. L. Harper & B. Skyrms (Eds.), *Causation in decision, belief change, and statistics* (Vol. II, pp. 105–134). Dordrecht/Boston: Kluwer Academic

Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (APQ monograph series, Vol. 2). Oxford: Blackwell.

Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy, 12*, 133–163.

Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies, 128*(1), 169–199.

van Benthem, J. F. A. K. (2007). Dynamic logic for belief revision. *Journal of Applied Non-classical Logics, 17*(2), 129–155.

van Benthem, J. F. A. K., Gerbrandy, J., & Kooi, B. (2006a) Dynamic update with probabilities. In Bonanno et al. (2006) (pp. 237–246).

van Benthem, J. F. A. K., & Liu, F. (2004). Dynamic logic of preference upgrade. Technical report, University of Amsterdam. ILLC Publications, PP-2005-29.

van Benthem, J. F. A. K., van Eijck, J., & Kooi, B. P. (2006b). Logics of communication and change. *Information and Computation, 204*(11), 1620–1662.

van der Hoek, W. (1993). Systems for knowledge and beliefs. *Journal of Logic and Computation, 3*(2), 173–195.

van Ditmarsch, H. P. (2000). *Knowledge games*. PhD thesis, University of Groningen. ILLC Pubications, DS-2000-06.

van Ditmarsch, H. P. (2002). Descriptions of game actions. *Journal of Logic, Language and Information, 11*(3), 349–365.

van Ditmarsch, H. P. (2005) Prolegomena to dynamic logic for belief revision. *Synthese, 147*(2), 229–275.

van Ditmarsch, H. P., & Labuschagne, W. (2007). My beliefs about your beliefs: A case study in theory of mind and epistemic logic. *Synthese, 155*(2), 191–209.

van Ditmarsch, H. P., van der Hoek, W., & Kooi, B. P. (2007). *Dynamic epistemic logic* (Synthese library, Vol. 337). Dordrecht: Springer.

Voorbraak, F. P. J. M. (1993). *As far as I know*. PhD thesis, Utrecht University, Utrecht (Quaestiones infinitae, Vol. VII).

Williamson, T. (2001). Some philosophical aspects of reasoning about knowledge. In J. van Benthem (Ed.), *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge (TARK'01)* (p. 97). San Francisco: Morgan Kaufmann.

Chapter 40
Agreeing to Disagree

Robert J. Aumann

If two people have the same priors, and their posteriors for a given event A are common knowledge, then these posteriors must be equal. This is so even though they may base their posteriors on quite different information. In brief, people with the same priors *cannot agree to disagree*.

We publish this observation with some diffidence, since once one has the appropriate framework, it is mathematically trivial. Intuitively, though, it is not quite obvious; and it is of some interest in areas in which people's beliefs about each other's beliefs are of importance, such as game theory[1] and the economics of information.[2] A "concrete" illustration that may clarify matters (and that may be read at this point) is found at the end of the paper.

The key notion is that of "common knowledge." Call the two people 1 and 2. When we say that an event is "common knowledge," we mean more than just that both 1 and 2 know it; we require also that 1 knows that 2 knows it, 2 knows that 1 knows it, 1 knows that 2 knows that 1 knows it, and so on. For example, if 1 and 2 are both present when the event happens and see each other there, then the event

AMS 1970 *subject classifications.* Primary 62A15, 62C05; Secondary 90A05, 90D35.

Key words and phrases. Information, subjective probability, posterior, statistics, game theory, revising probabilities, concensus, Harsanyi doctrine.

[1]Cf. Harsanyi (1967–1968); also Aumann (1974), especially Section 9j (page 92), in which the question answered here was originally raised.

[2]Cf., e.g., Radner (1968, 1972); also the review by Grossman and Stiglitz (1976) and the papers quoted there.

R.J. Aumann (✉)
Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: raumann@math.huji.ac.il

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_40

becomes common knowledge. In our case, if 1 and 2 tell each other their posteriors and trust each other, then the posteriors are common knowledge. The result is not true if we merely assume that the persons know each other's posteriors.

Formally, let $(\Omega, \mathscr{B}, p)$ be a probability space, $\mathscr{P}_1$ and $\mathscr{P}_2$ partitions of Ω whose join[3] $\mathscr{P}_1 \vee \mathscr{P}_2$ consists of nonnull events.[4] In the interpretation, $(\Omega, \mathscr{B})$ is the space of states of the world, p the common prior of 1 and 2, and $\mathscr{P}_i$ the information partition of i; that is, if the true state of the world is ω, then i is informed of that element $\mathbf{P}_i(\omega)$ of $\mathscr{P}_i$ that contains ω. Given ω in Ω, an event E is called *common knowledge at* ω if E includes that member of the meet[5] $\mathscr{P}_1 \wedge \mathscr{P}_2$ that contains ω. We will show below that this definition is equivalent to the informal description given above.

Let A be an event, and let $\mathbf{q}_i$ denote the posterior probability $p(A \mid \mathscr{P}_i)$ of A given i's information; i.e., if $\omega \in \Omega$, then $\mathbf{q}_i(\omega) = p(A \cap \mathbf{P}_i(\omega))/p(\mathbf{P}_i(\omega))$.

Proposition *Let $\omega \in \Omega$, and let q_1 and q_2 be numbers. If it is common knowledge at ω that $\mathbf{q}_1 = q_1$ and $\mathbf{q}_2 = q_2$, then $q_1 = q_2$.*

Proof Let P be the member of $\mathscr{P}_1 \wedge \mathscr{P}_2$ that contains ω. Write $P = \cup_j P^j$, where the P^j are disjoint members of $\mathscr{P}_1$. Since $\mathbf{q}_1 = q_1$ throughout P, we have $p(A \cap P^j)/p(P^j) = q_1$ for all j; hence $p(A \cap P^j) = q_1 p(P^j)$, and so by summing over j we get $p(A \cap P) = q_1 p(P)$. Similarly $p(A \cap P) = q_2 p(P)$, and so $q_1 = q_2$. This completes the proof.

To see that the formal definition of "common knowledge" is equivalent to the informal description, let $\omega \in \Omega$, and call a member ω' of Ω *reachable from* ω if there is a sequence $P^1, P^2, \ldots, P^k$ such that $\omega \in P^1$, $\omega' \in P^k$, and consecutive P^j intersect and belong alternatively to $\mathscr{P}_1$ and $\mathscr{P}_2$. Suppose now that ω is the true state of the world, $P^1 = \mathbf{P}_1(\omega)$, and E is an event. To say that 1 "knows" E means that E includes P^1. To say that 1 knows that 2 knows E means that E includes all P^2 in $\mathscr{P}_2$ that intersect P^1. To say that 1 knows that 2 knows that 1 knows E means that E includes all P^3 in $\mathscr{P}_1$ that intersect P^2 in $\mathscr{P}_2$ that intersect P^1. And so on. Thus all sentences of the form "i knows that i' knows that i knows $\ldots$ E" (where $i' = 3 - i$) are true if and only if E contains all ω' reachable from ω. But the set of all ω' reachable from ω is a member of $\mathscr{P}_i \wedge \mathscr{P}_2$; so the desired equivalence is established.

The result fails when people merely know each other's posteriors. Suppose Ω has 4 elements $\alpha, \beta, \gamma, \delta$ of equal (prior) probability, $\mathscr{P}_1 = \{\alpha\beta, \gamma\delta\}$, $\mathscr{P}_2 = \{\alpha\beta\gamma, \delta\}$, $A = \alpha\delta$, and $\omega = \alpha$. Then 1 knows that $\mathbf{q}_2$ is $\frac{1}{3}$, and 2 knows that $\mathbf{q}_1$ is $\frac{1}{2}$; but 2 thinks that 1 may not know what $\mathbf{q}_2$ is ($\frac{1}{3}$ or 1).

Worthy of note is the implicit assumption that the information partitions $\mathscr{P}_1$ and $\mathscr{P}_2$ are themselves common knowledge. Actually, this constitutes no loss of

[3] Coarsest common refinement of $\mathscr{P}_1$ and $\mathscr{P}_2$.

[4] Events whose (prior) probability does not vanish.

[5] Finest common coarsening of $\mathscr{P}_1$ and $\mathscr{P}_2$.

generality. Included in the full description of a state ω of the world is the manner in which information is imparted to the two persons. This implies that the information sets $\mathbf{P}_1(\omega)$ and $\mathbf{P}_2(\omega)$ are indeed defined unambiguously as functions of ω, and that these functions are known to both players.

Consider next the assumption of equal priors for different people. John Harsanyi (1968) has argued eloquently that differences in subjective probabilities should be traced exclusively to differences in information—that there is no rational basis for people who have always been fed precisely the same information to maintain different subjective probabilities. This, of course, is equivalent to the assumption of equal priors. The result of this paper might be considered evidence against this view, as there are in fact people who respect each other's opinions and nevertheless disagree heartily about subjective probabilities. But this evidence is not conclusive: even people who respect each other's acumen may ascribe to each other errors in calculating posteriors. Of course we do not mean simple arithmetical mistakes, but rather systematic biases such as those discussed by Tversky and Kahneman (1974). In private conversation, Tversky has suggested that people may also be biased because of psychological factors that may make them disregard information that is unpleasant or does not conform to previously formed notions.

There is a considerable literature about reaching agreement on subjective probabilities; a recent paper is DeGroot (1974), where a bibliography on the subject may be found. A "practical" method is the Delphi technique (see, e.g., Dalkey (1972)). It seems to me that the Harsanyi doctrine is implicit in much of this literature; reconciling subjective probabilities makes sense if it is a question of implicitly exchanging information, but not if we are talking about "innate" differences in priors. The result of this paper might be considered a theoretical foundation for the reconciliation of subjective probabilities.

As an illustration, suppose 1 and 2 have a uniform prior on the parameter of a coin, and let A be the event that the coin will come up H (heads) on the next toss. Suppose that each person is permitted to make one previous toss, and that these tosses come up H and T (tails) respectively. If each one's information consists precisely of the outcome of his toss, then the posteriors for A will be $\frac{2}{3}$ and $\frac{1}{3}$ respectively. If each one then informs the other one of his posterior, then they will both conclude that the previous tosses came up once H and once T, so that both posteriors will be revised to $\frac{1}{2}$.

Suppose now that each person is permitted to make several previous tosses, but that neither one knows how many tosses are allowed to the other one. For example, perhaps both make 4 tosses, which come up *HHHT* for 1, and *HTTT* for 2. They then inform each other that their posteriors are $\frac{2}{3}$ and $\frac{1}{3}$ respectively. Now these posteriors may result from a single observation, from 4 observations, or from more. Since neither one knows on what observations the other's posterior is based, he may be inclined to give more weight to his own observations. Some revision of posteriors would certainly be called for even in such a case; but it does not seem clear that it would necessarily lead to equal posteriors.

Presumably, such a revision would take into account each person's prior on the number of tosses available to him and to the other person. By assumption these two

priors are the same, but each person gets additional private information—namely, the actual number of tosses he is allotted. By use of the prior and the information that the posteriors are, respectively, $\frac{2}{3}$ and $\frac{1}{3}$, new posteriors may be calculated. If the players inform each other of these new posteriors, further revision may be called for. Our result implies that the process of exchanging information on the posteriors for A will continue until these posteriors are equal.

Acknowledgments This work was supported by National Science Foundation Grant SOC74-11446 at the Institute for Mathematical Studies in the Social Sciences, Stanford University.

References

Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics, 1*, 67–96.

Dalkey, N. C. (1972). *Studies in the quality of life*. Lexington: Lexington Books.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association, 69*, 118–121.

Grossman, S. J., & Stiglitz, J. E. (1976). Information and competitive price systems. *The American Economic Review, 46*, 246–253.

Harsanyi, J. (1967–1968). Games of incomplete information played by Bayesian players, Parts I–III. *Management Science, 14*, 159–182, 320–334, 486–502.

Radner, R. (1968). Competitive equilibrium under uncertainty. *Econometrica, 36*, 21–58.

Radner, R. (1972). Existence of equilibrium plans, prices, and price expectations in a sequence of markets. *Econometrica, 40*, 289–304.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Chapter 41
Epistemic Conditions for Nash Equilibrium

Robert J. Aumann and Adam Brandenburger

Introduction

Game theoretic reasoning has been widely applied in economics in recent years. Undoubtedly, the most commonly used tool has been the strategic equilibrium of Nash (1951), or one or another of its so-called "refinements." Though much effort[1] has gone into developing these refinements, relatively little attention has been paid to a more basic question: Why consider Nash equilibrium in the first place?

A Nash equilibrium is defined as a way of playing the game—an n-tuple of strategies—in which each player's strategy is optimal for him, given the strategies of the others. The definition seems beautifully simple and compelling; but when considered carefully, it lacks a clear motivation. What would make the players play such an equilibrium? How, exactly, would it come about?

Over the years, much has been written about the connection between Nash equilibrium and common knowledge.[2] According to conventional wisdom, Nash

[1] Selten (1965), (1975), Myerson (1978), Kreps and Wilson (1982), Kohlberg and Mertens (1986), and many others.

[2] An event is called *common knowledge* if all players know it, all know that all know it, and so on ad infinitum (Lewis 1969).

R.J. Aumann (✉)
Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: raumann@math.huji.ac.il

A. Brandenburger
Stern School of Business, Tandon School of Engineering, NYU Shanghai, New York University, New York, NY 10012, USA
e-mail: adam.brandenburger@stern.nyu.edu; adambrandenburger.com

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_41

equilibrium is "based on" common knowledge of (a) the structure of the game (i.e., the payoff functions[3]), (b) the rationality[4] of the players, and (c) the strategies actually played. These ideas sound appealing; the circularity of Nash equilibrium—each player chooses his strategy only because the others choose theirs—does seem related to the infinite hierarchy of beliefs inherent in common knowledge. But a formalization has proved elusive. What, precisely, does "based on" mean? Can the above wisdom be turned into a theorem?

It is our purpose here to clarify these issues in a formal framework. Specifically, we seek *epistemic* conditions for a given strategy profile to be a Nash equilibrium: conditions involving what the players know or believe about one another—in particular, about payoff functions, strategy choices, decision procedures, and beliefs about these matters.[5]

Surprisingly, we will find that common knowledge of the payoff functions and of rationality are *never* needed; much weaker epistemic conditions suffice. Common knowledge of the strategies actually being played is also irrelevant. What *does* turn out to be relevant is common knowledge of the beliefs that the players hold about the strategies of the others; but here, too, common knowledge is relevant only when there are at least three players.

The main results are informally described in section "Description of the results." The background for a formal presentation is given in section "Interactive belief systems"; the underlying tool is that of an *interactive belief system*, which provides a framework for formulating epistemic conditions. Illustrations of such systems are given in section "Illustrations." The formal statements and proofs of the main results, in section "Formal statements and proofs of the theorems," are preceded by some lemmas in section "Properties of belief systems." Sections "The main counterexamples" and "Additional counterexamples" contain a series of counterexamples showing the results to be sharp; the examples—particularly those of section "The main counterexamples"—provide insight into the role played by the various epistemic conditions. Section "General (infinite) belief systems" shows that the results apply to infinite as well as finite belief systems. Section "Discussion" is devoted to a discussion of conceptual aspects and of the related literature. An appendix treats extensions and converses of the main results.

The reader wishing to understand just the main ideas should read sections "Description of the results" and "The main counterexamples," and skim sections "Interactive belief systems" and "Illustrations."

[3] See section "Structure of the game" for a discussion of why the payoff functions can be identified with the "structure of the game."

[4] I.e., that the players are optimizers; that given the opportunity, they will choose a higher payoff. A formal definition is given below.

[5] Other epistemic conditions for Nash equilibrium have been obtained by Armbruster and Boege (1979) and Tan and Werlang (1988).

Description of the Results

An event is called *mutual knowledge* if all players simply know it (to be distinguished from common knowledge, which also requires higher knowledge levels—knowledge about knowledge, and so on). Our first and simplest result is Theorem 41.1: *Suppose that each player is rational and knows his own payoff function, and that the strategy choices of the players are mutually known. Then these choices constitute a Nash equilibrium in the game being played.*

The proof is immediate: Since each player knows the choices of the others, and is rational, his choice must be optimal given theirs; so by definition, we are at a Nash equilibrium.

Note that neither the players' rationality, nor their payoff functions, nor their strategy choices are assumed common knowledge. For strategies, only mutual knowledge is assumed. For rationality and the structure of the game, not even mutual knowledge is assumed; only that the players are in fact rational, and know their own payoff functions.[6]

Theorem 41.1 applies to all pure strategy profiles. It applies also to mixed strategy profiles, under the traditional view of mixed strategies as conscious randomizations; in that case, of course, it is the mixture that must be mutually known, not just their pure realizations.

In recent years, a different view of mixed strategies has emerged.[7] In this view, players do not randomize; each player chooses some definite pure strategy. But the other players need not know which one, and the mixture represents their uncertainty, their probability assessment of his choice. This is the view adopted in the sequel; it fits in well with the Bayesian approach to game theory, in which uncertainty about strategic choices of others is, like any other uncertainty, subject to probability assessment by each player.

For brevity, let us refer to pure strategies as *actions*. Define the *conjecture* of a player as his probability assessment of the actions of the other players. Call a player *rational* if his action maximizes his expected payoff given his conjecture.

When there are two players (and only then), the conjecture of each player is a mixed strategy of the other player. Because of this, the results in the two-person and *n*-person cases are quite different. For *two*-person games, we have the following (Theorem 41.2): *Suppose that the game being played (i.e., both payoff functions), the rationality of the players, and their conjectures are all mutually known. Then the conjectures constitute a Nash equilibrium.*

Theorem 41.2 differs from Theorem 41.1 in two ways. First, in both the conclusion and the hypothesis, strategy choices are replaced by conjectures; thus we get "conjectural equilibrium"—an equilibrium in conjectures, not in strategies

[6] When a game is presented in strategic form, as here, knowledge of one's own payoff function may be considered tautologous. See section "Interactive belief systems."

[7] Harsanyi (1973), Armbruster and Boege (1979), Aumann (1987), Tan and Werlang (1988), Brandenburger and Dekel (1989), among others.

actually played. Second, the hypothesis calls not just for the fact of rationality, but for mutual knowledge of this fact, and for mutual knowledge of the payoff functions. But common knowledge still does not enter the picture.

Since we are now viewing mixed strategies as conjectures, it is natural that conjectures replace choices in the result. So with n players, too, one might expect a theorem roughly analogous to Theorem 41.2; i.e., that mutual knowledge of the conjectures (when combined with appropriate assumptions about rationality and the payoff functions) is sufficient for them to be in equilibrium. But here we are in for a surprise: when $n > 2$, the conditions for a conjectural equilibrium become much more stringent.

To understand the situation, note that the conjecture of each player i is a probability mixture of $(n-1)$-tuples of pure strategies of the other players. So when $n > 2$, it is not itself a mixed strategy; however, it induces a mixed strategy[8] for each player j other than i, called *i's conjecture about j*. One difficulty is that different players other than j may have different conjectures about j, in which case it is not clear how to define j's component of the conjectural equilibrium we seek to construct.

To present Theorem 41.3, the n-person "conjecture theorem", one more concept is needed. We say that the players have a *common prior*[9] if all differences between their probability assessments are due only to differences in their information; more precisely, if there is an outside observer O with no private information,[10] such that for all players i, if O were given i's information, his probability assessments would be the same as i's.

Theorem 41.3 is now as follows: *In an n-player game, suppose that the players have a common prior, that their payoff functions and their rationality are mutually known, and that their conjectures are commonly known. Then for each player j, all the players i agree on the same conjecture σ_j about j, and the resulting profile $(\sigma_1, \ldots, \sigma_n)$ of mixed strategies is a Nash equilibrium.*

The above three theorems give sufficient epistemic conditions for Nash equilibrium. The conditions are not necessary; it is always possible for the players to blunder into a Nash equilibrium "by accident," so to speak, without anybody knowing much of anything. Nevertheless, all three theorems are "sharp," in the sense that they cannot be improved upon; none of the conditions can be dispensed with, or, so far as we can see, significantly weakened.

The presentation in this section, while correct, has been informal. For a formal presentation, one needs a framework for describing "epistemic" situations in game contexts; in which, for example, one can describe a situation where each player maximizes against the choices of the others, all know this, but not all know that all know this. Such frameworks are available in the differential information literature; a particular adaptation is presented in the next section.

[8] The marginal on j's strategy space of i's overall conjecture.

[9] Aumann (1987); for a formal definition, see section "Interactive belief systems." Harsanyi (1967–1968) uses the term "consistency" to describe this situation.

[10] That is, what O knows is common knowledge among the players; each player knows everything that O knows.

Interactive Belief Systems

Let us be given a strategic *game form*; that is, a finite set $\{1, \ldots, n\}$ (the *players*), together with an action set A_i for each player i. Set $A := A_1 \times \cdots \times A_n$. An *interactive belief system* (or simply *belief system*) for this game form is defined to consist of:

(1) for each player i, a set S_i (i's *types*),
and for each type s_i of i,
(2) a probability distribution on the set S^{-i} of $(n-1)$-tuples of types of the other players (s_i's *theory*),
(3) an action a_i for i (s_i's *action*),
and
(4) a function $g_i : A \to \mathbb{R}$ (s_i's *payoff function*).

The action sets A_i are assumed finite. One may also think of the spaces S_i as finite; the ideas are then more transparent. For a general definition, where the S_i are measurable spaces and the theories are probability measures,[11] see section "General (infinite) belief systems."

A belief system is a formal description of the players' beliefs—about each other's actions and payoff functions, about these beliefs, and so on. Specifically, the theory of a type s_i represents the probabilities that s_i ascribes to the types of the other players, and so to their actions, their payoff functions, and their theories. See section "Belief systems" for further discussion.

Set $S := S_1 \times \cdots \times S_n$. Call the members $s = (s_1, \ldots, s_n)$ of S *states of the world*, or simply *states*. An *event* is a subset E of S. Denote by $p(\cdot; s_i)$ the probability distribution on S induced by s_i's theory; formally, if E is an event, then $p(E; s_i)$ is the probability assigned by s_i to $\{s^{-i} \in S^{-i} : (s_i, s^{-i}) \in E\}$.

A function $g : A \to \mathbb{R}^n$ (an n-tuple of payoff functions) is called a *game*.

Set $A^{-i} := A_1 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_n$; for a in A set $a^{-i} := (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$. When referring to player i, the phrase "at s" means "at s_i". Thus, "i's action at s" means s_i's action (see (3)); we denote it $\mathbf{a}_i(s)$, and write $\mathbf{a}(s)$ for the n-tuple $(\mathbf{a}_1(s), \ldots, \mathbf{a}_n(s))$ of actions at s. Similarly, "i's payoff function at s" means s_i's payoff function (see (4)); we denote it $\mathbf{g}_i(s)$, and write $\mathbf{g}(s)$ for the n-tuple $(\mathbf{g}_1(s), \ldots, \mathbf{g}_n(s))$ of payoff functions[12] at s. Viewed as a function of a, we call $\mathbf{g}(s)$ "the game being played at s," or simply "the game at s."

[11] Readers unfamiliar with measure theory may think of the type spaces S_i as finite throughout the paper. All the examples involve finite S_i only. The results, too, are stated and proved without reference to measure theory, and may be understood completely in terms of finite S_i. On the other hand, we do not *require* finite S_i; the definitions, theorems, and proofs are all worded so that when interpreted as in section "General (infinite) belief systems," they apply without change to the general case. One can also dispense with finiteness of the action spaces A_i; but that is both more involved and less important, and we will not do it here.

[12] Thus i's actual payoff at the state s is $\mathbf{g}_i(s)(\mathbf{a}(s))$.

Functions defined on S (like $\mathbf{a}_i(s)$, $\mathbf{a}(s)$, $\mathbf{g}_i(s)$, and $\mathbf{g}(s)$) may be viewed like random variables in probability theory. Thus if $\mathbf{x}$ is such a function and x is one of its values, then $[\mathbf{x} = x]$, or simply $[x]$, denotes the event $\{s \in S : \mathbf{x}(s) = x\}$. For example, $[a_i]$ denotes the event that i chooses the action a_i; and $[g]$ denotes the event that the game g is being played.

A *conjecture* φ^i of i is a probability distribution on A^{-i}. For $j \neq i$, the marginal of φ^i on A_j is called the *conjecture of i about j induced by φ^i*. The theory of i at a state s yields a conjecture $\boldsymbol{\varphi}^i(s)$, called *$i$'s conjecture at s*, given by $\boldsymbol{\varphi}^i(s)\,(a^{-i}) := p\,([a^{-i}]; s_i)$. We denote the n-tuple $(\boldsymbol{\varphi}^1(s), \ldots, \boldsymbol{\varphi}^n(s))$ of conjectures at s by $\boldsymbol{\varphi}(s)$.

Player i is called *rational at s* if his action at s maximizes his expected payoff given his information (i.e., his type s_i); formally, letting $h_i := \mathbf{g}_i(s)$ and $b_i, := \mathbf{a}_i(s)$, this means that $\mathrm{Exp}\,(h_i\,(b_i, \mathbf{a}^{-i})\,|s_i) \geq \mathrm{Exp}\,(h_i\,(a_i, \mathbf{a}^{-i})\,|s_i)$ for all a^i in A^i. Another way of saying this is that i's actual choice b_i maximizes the expectation of his actual payoff h_i when the other players' actions are distributed according to his actual conjecture $\boldsymbol{\varphi}^i(s)$.

Player i is said to *know* an event E *at s* if at s, he ascribes probability 1 to E. Define K_iE as the set of all those s at which i knows E. Set $K^1E := K_1E \cap \cdots \cap K_nE$; thus K^1E is the event that all players know E. If $s \in K^1E$, call E *mutually known at s*. Set $CKE := K^1E \cap K^1K^1E \cap K^1K^1K^1E \cap \cdots$; if $s \in CKE$, call E *commonly known at s*.

A probability distribution P on S is called a *common prior* if for all players i and all of their types s_i, the conditional distribution of P given s_i is $p(\cdot; s_i)$; this implies that for all i, all events E and F, and all numbers π,

(5) if $p(E; s_i) = \pi\ p(F; s_i)$ for all $s_i \in S_i$, then $P(E) = \pi P(F)$.

In words, (5) says that for each player i, if two events have proportional probabilities given any s_i, then they have proportional prior probabilities.

Another regularity condition that is sometimes used in the differential information literature is "mutual absolute continuity." We do not define this here because we have no use for it.

Belief systems provide a formal language for stating epistemic conditions. When we say that a player knows some event E, or is rational, or has a certain conjecture φ^i or payoff function g_i, we mean that that is the case at some specific state s of the world. Thus at s, Rowena may know E, but not know that Colin knows E; or at s, it may be that Colin is rational, and that Rowena knows this, but that Colin does not know that Rowena knows it. Some illustrations of these ideas are given in the next section.

Illustrations

Example 41.1 We start with a belief system in which all types of each player i have the same payoff function g_i, namely, that depicted in Fig. 41.1. Thus the game being played is commonly known. Call the row and column players (Players 1 and 2) "Rowena" and "Colin" respectively.

Fig. 41.1

	c	d
C	2, 2	0, 0
D	0, 0	1, 1

Fig. 41.2

	c_1	d_1	d_2
C_1	2/5, 2/3	3/5, 3/5	0, 0
D_1	1/2, 1/3	0, 0	1/2, 1/2
D_2	0, 0	2/3, 2/5	1/3, 1/2

The theories are depicted in Fig. 41.2; here C_1 denotes a type of Rowena whose action is C, whereas D_1 and D_2 denote two different types of Rowena whose actions are D. Similarly for Colin. Each square denotes a state, i.e., a pair of types. The two entries in each square denote the probabilities that the corresponding types of Rowena and Colin ascribe to that state. For example, Colin's type d_2 attributes $1/2 - 1/2$ probabilities to Rowena's type being D_1 or D_2. So at state (D_2, d_2), he knows that Rowena will choose the action D. Similarly, Rowena knows at (D_2, d_2) that Colin will choose d. Since d and D are optimal against each other, both players are rational at (D_2, d_2), and (D, d) is a Nash equilibrium.

We have here a typical instance of Theorem 41.1 (see the beginning of section "Description of the results"), which also shows that the folk wisdom cited in the introduction is misleading. At (D_2, d_2), there is mutual knowledge of the actions D and d, and both players are in fact rational. But the actions are not common knowledge. Thus, though Colin knows that Rowena will play D, she doesn't know that he knows this; indeed, she attributes probability 2/3 to his attributing probability 3/5 to her playing C. Moreover, though both players are rational at (D_2, d_2), there isn't even mutual knowledge of rationality there. For example, Colin's type d_1 chooses d, with an expected payoff of 2/5, rather than c, with an expected payoff of 6/5; thus this type is "irrational." At (D_2, d_2), Rowena attributes probability 2/3 to Colin being of this irrational type.

Note that the players have a common prior (Fig. 41.3). But Theorem 41.1 has nothing to do with common priors. If, for example, we change the theory of Rowena's type D_2 from $\frac{2}{3}d_1 + \frac{1}{3}d_2$ to $\frac{1}{2}d_1 + \frac{1}{2}d_2$, then there is no longer a common prior; but (D, d) is still Nash equilibrium, for the same reasons as above. (As usual,

Fig. 41.3

	c_1	d_1	d_2
C_1	0.2	0.3	0
D_1	0.1	0	0.1
D_2	0	0.2	0.1

Fig. 41.4

	h	t
H	1, 0	0, 1
T	0, 1	1, 0

Fig. 41.5

	h_1	t_1	t_2
H_1	1/2, 1/2	1/2, 1/2	0, 0
T_1	1/2, 1/2	0, 0	1/2, 1
T_2	0, 0	1, 1/2	0, 0

$\frac{2}{3}d_1 + \frac{1}{3}d_2$ stands for the $\frac{2}{3} - \frac{1}{3}$ probability combination of d_1 and d_2. Similar notation will be used throughout the sequel.)

Example 41.2 As in the previous example, the game being played here—"matching pennies" (Fig. 41.4)—is commonly known. The theories are depicted in Fig. 41.5. At the state (H_1, h_1), the conjectures of Rowena and Colin are $\frac{1}{2}h + \frac{1}{2}t$ and $\frac{1}{2}H + \frac{1}{2}T$ respectively, and these conjectures are mutually known (i.e., each knows that these are the conjectures). Moreover, the rationality of both players is mutually known. Thus Theorem 41.2 (section "Description of the results") implies that $(\frac{1}{2}H + \frac{1}{2}T, \frac{1}{2}h + \frac{1}{2}t)$ is a Nash equilibrium, which indeed it is.

Note that neither the conjectures of the players nor their rationality are commonly known. Indeed, at (T_1, t_2), Colin knows that Rowena plays T, so that his conjecture is not $\frac{1}{2}H + \frac{1}{2}T$ but T; so it is irrational for him to play t, which yields him 0, rather than 1 that he could get by playing h. At the state (H_1, h_1), Colin attributes probability 1/2 to Rowena attributing probability 1/2 to the state (T_1, t_2); so at (H_1, h_1), there is common knowledge neither of Colin's conjecture $\frac{1}{2}H + \frac{1}{2}T$, nor of his rationality.

Fig. 41.6

	h_1	t_1	t_2
H_1	0.2	0.2	0
T_1	0.2	0	0.2
T_2	0	0.2	0

Note, too, that like in the previous example, this belief system has a common prior (Fig. 41.6). But also like there, this is not essential; the discussion would not be affected if, say, we changed the theory of Rowena's type T_2 from t_1 to $\frac{1}{2}t_1 + \frac{1}{2}t_2$.

Properties of Belief Systems

In this section we formally establish some basic properties of belief systems, which are needed in the sequel. These properties are intuitively fairly obvious, and some are well known in various formalizations of interactive knowledge theory; so this section can be omitted at a first reading.

Lemma 41.1 *Player i knows that he attributes probability π to an event E if and only if he indeed attributes probability π to E.*

Proof If: Let F be the event that i attributes probability π to E; that is, $F := \{t \in S : p(E; t_i) = \pi\}$. If $s \in F$, then $p(E; s_i) = \pi$, so all states u with $u_i = s_i$ are in F. Therefore $p(F; s_i) = 1$; that is, i knows F at s, so $s \in K_i F$.

Only if: Suppose that i attributes probability $\rho \neq \pi$ to E. By the "if" part of the proof, he must know this, contrary to his knowing that he attributes probability π to E. ■

Corollary 41.1 *Let φ be an n-tuple of conjectures. Suppose that at some state s, it is mutually known that $\boldsymbol{\varphi} = \varphi$. Then $\boldsymbol{\varphi}(s) = \varphi$. (In words: if it is mutually known that the conjectures are φ, then they are indeed φ.)*

Corollary 41.2 *A player is rational if and only if he knows that he is rational.*

Corollary 41.3 *$K_i K_i E = K_i E$ (a player knows something if and only if he knows that he knows it), and $K_i \neg K_i E = \neg K_i E$ (a player doesn't know something if and only if he knows that he doesn't know it).*

Lemma 41.2 *$K_i(E_1 \cap E_2 \cap \cdots) = K_i E_1 \cap K_i E_2 \cap \cdots$ (a player knows each of several events if and only if he knows that they all obtain).*

Proof At s, player i ascribes probability 1 to $E_1 \cap E_2 \cap \cdots$ if and only if he ascribes probability 1 to each $E_1, E_2, \ldots$. ■

Lemma 41.3 *$CKE \subset K_i CKE$ (if something is commonly known, then each player knows that it is commonly known).*

Proof Since $K_i K^1 F \supset K^1 K^1 F$ for all F, Lemma 41.2 yields $K_i CKE = K_i(K^1 E \cap K^1 K^1 E \cap \cdots) = K_i K^1 E \cap K_i K^1 K^1 E \cap \cdots \supset K^1 K^1 E \cap K^1 K^1 K^1 E \cap \cdots \supset CKE$. ■

Lemma 41.4 *Suppose P is a common prior, $K_i H \supset H$, and $p\,(E; s_i) = \pi$ for all $s \in H$. Then $P(E \cap H) = \pi P(H)$.*

Proof Let H_i be the projection of H on S_i. From $K_i H \supset H$ it follows that $p\,(H; s_i) = 1$ or 0 according to whether s_i is or is not[13] in H_i. So when $s_i \in H_i$, then $p(E \cap H; s_i) = p\,(E; s_i) = \pi = \pi p\,(H; s_i)$; and when $s_i \notin H_i$, then $p(E \cap H; s_i) = 0 = \pi p(H; s_i)$. The lemma now follows from (5). ■

The following lemma is not needed for the proofs of the theorems, but relates to the examples in section "The main counterexamples." Set $K^2 E := K^1 K^1 E$, $K^3 E := K^1 K^2 E$, and so on. If $s \in K^m E$, call E *mutually known of order m* at s.

Lemma 41.5 *$K^m E \subset K^{m-1} E$ for $m > 1$ (mutual knowledge of order m implies mutual knowledge of orders $1, \ldots, m-1$).*

Proof By Lemma 41.2 and Corollary 41.3, $K^2 E = K^1 K^1 E = \cap_{i=1}^n K_i K^1 E \subset K_i K^1 E = K_i \cap_{j=1}^n K_j E \subset K_i K_i E$. Since this is so for all i, we get $K^2 E \subset \cap_{i=1}^n K_i E = K^1 E$. The result for $m > 2$ follows from this by substituting $K^{m-2} E$ for E. ■

Formal Statements and Proofs of the Theorems

We now formally state and prove Theorems 41.1, 41.2, and 41.3. For more transparent paraphrases (using the same terminology), see section "Description of the results."

Theorem 41.1 *Let $\mathbf{a}$ be an n-tuple of actions. Suppose that at some state s, all players are rational, and all know that $\mathbf{a} = a$. Then a is a Nash equilibrium.*

Proof Immediate (see beginning of section "Description of the results"). ■

Theorem 41.2 *With $n = 2$ (two players), let g be a game, φ a pair of conjectures. Suppose that at some state, it is mutually known that $\mathbf{g} = g$, that the players are rational, and that $\boldsymbol{\varphi} = \varphi$. Then (φ^2, φ^1) is a Nash equilibrium of g.*

The proof uses a lemma; we state it for the n-person case, since it is needed again in the proof of Theorem 41.3.

Lemma 41.6 *Let g be a game, φ an n-tuple of conjectures. Suppose that at some state s, it is mutually known that $\mathbf{g} = g$, that the players are rational, and that $\boldsymbol{\varphi} = \varphi$.*

[13] In particular, i always knows whether or not H obtains.

Let a_j be an action of a player j to which the conjecture φ^i of some other player i assigns positive probability. Then a_j maximizes g_j against[14] φ^j.

Proof By Corollary 41.1, the conjecture of i at s is φ^i. So i attributes positive probability at s to $[a_j]$. Also, i attributes probability 1 at s to each of the three events [j is rational], $[\varphi^j]$, and $[g_j]$. When one of four events has positive probability, and the other three each have probability 1, then their intersection is non-empty. So there is a state t at which all four events obtain: j is rational, he chooses a_j, his conjecture is φ^j, and his payoff function is g_j. So a_j maximizes g_j against φ^j. ■

Proof of Theorem 41.2 By Lemma 41.6, every action a_1 with positive probability in φ^2 is optimal against φ^1 in g, and every action a_2 with positive probability in φ^1 is optimal against φ^2 in g. This implies that (φ^2, φ^1) is a Nash equilibrium of g. ■

Theorem 41.3 *Let g be a game, φ an n-tuple of conjectures. Suppose that the players have a common prior, which assigns positive probability to it being mutually known that $\mathbf{g} = g$, mutually known that all players are rational, and commonly known that $\boldsymbol{\varphi} = \varphi$. Then for each j, all the conjectures φ^i of players i other than j induce the same conjecture σ_j for j, and $(\sigma_1, \ldots, \sigma_n)$ is a Nash equilibrium of g.*

The proof requires a lemma.

Lemma 41.7 *Let Q be a probability distribution on A with*[15] $Q(a) = Q(a_i)\, Q(a^{-i})$ *for all a in A and all i. Then $Q(a) = Q(a_1) \cdots Q(a_n)$ for all a.*

Proof By induction. For $n = 1$ and 2 the result is immediate. Suppose it true for $n-1$. From $Q(a) = Q(a_1)Q(a^{-1})$ we obtain, by summing over a_n, that $Q(a^{-n}) = Q(a_1)\, Q(a_2, \ldots, a_{n-1})$. Similarly $Q(a^{-n}) = Q(a_1)\, Q(a_i, \ldots, a_{i-1}, a_{i+1}, \ldots, a_{n-1})$ whenever $i < n$. So the induction hypothesis yields $Q(a^{-n}) = Q(a_1)Q(a_2) \cdots Q(a_{n-1})$. Hence $Q(a) = Q(a^{-n})\, Q(a_n) = Q(a_1)\, Q(a_2) \cdots Q(a_n)$. ■

Proof of Theorem 41.3 Set $F := CK[\varphi]$, and let P be the common prior. By assumption, $P(F) > 0$. Set $Q(a) := P([a] \mid F)$. We show that for all a and i,

$$Q(a) = Q(a_i)\, Q\left(a^{-i}\right). \tag{41.1}$$

Set $H := [a_i] \cap F$. By Lemmas 41.2 and 41.3, $K_i H \supset H$, since i knows his own action. If $s \in H$, it is commonly, and so mutually, known at s that $\boldsymbol{\varphi} = \varphi$; so by Corollary 41.1, $\boldsymbol{\varphi}(s) = \varphi$; that is, $p([a^{-i}]; s_i) = \varphi^i(a^{-i})$. So Lemma 41.4 (with $E = [a^{-i}]$) yields $P([a] \mid F) = P([a^{-i}] \cap H) = \varphi^i(a^{-i})\, P(H) = \varphi^i(a^{-i})\, P([a_i] \mid F)$. Dividing by $P(F)$ yields $Q(a) = \varphi^i(a^{-i})\, Q(a_i)$; then summing over a_i, we get

$$Q\left(a^{-i}\right) = \varphi_i\left(a^{-i}\right). \tag{41.2}$$

Thus $Q(a) = Q(a^{-i})\, Q(a_i)$, which is (41.1).

[14] That is, $\operatorname{Exp} g_j(a_j, a^{-j}) \geq \operatorname{Exp} g_j(b_j, a^{-j})$ for all b_j in A_j, when a^{-j} is distributed according to φ^j.

[15] We denote $Q(a^{-i}) := Q(A_i \times \{a^{-i}\})$, $Q(a_i) := Q(A^{-i} \times \{a_i\})$, and so on.

For each j, define a probability distribution σ_j on A_j by $\sigma_j(a_j) := Q(a_j)$. Then (41.2) yields $\varphi^i(a_j) = Q(a_j) = \sigma_j(a_j)$ for $j \neq i$. Thus for all i, the conjecture for j induced by φ^i is σ_j, which does not depend on i. Lemma 41.7, (41.1), and (41.2) then yield

$$\varphi^i\left(a^{-i}\right) = \sigma_1(a_1) \cdots \sigma_{i-1}(a_{i-1})\, \sigma_{i+1}(a_{i+1}) \cdots \sigma_n(a_n); \tag{41.3}$$

that is, the distribution φ^i is the product of the distributions σ_j with $j \neq i$.

Since common knowledge implies mutual knowledge, the hypothesis of the theorem implies that there is a state at which it is mutually known that $\mathbf{g} = g$, that players are rational, and that $\boldsymbol{\varphi} = \varphi$. So by Lemma 41.6, each action a_j with $\varphi^i(a_j) > 0$ for some $i \neq j$ maximizes g_j against φ^j. By (41.3), these a_j are precisely the ones that appear with positive probability in σ_j. Again using (41.3), we conclude that each action appearing with positive probability in σ_j maximizes g_j against the product of the distributions σ_k with $k \neq j$. This implies that $(\sigma_1, \ldots, \sigma_n)$ is a Nash equilibrium of g.■

The Main Counterexamples

This section explores possible variations on Theorem 41.3 (the result giving sufficient epistemic conditions, when $n \geq 3$, for the players' conjectures to yield a Nash equilibrium). For simplicity, let $n = 3$. Each player's "overall" conjecture is then a distribution on pairs of actions of the other two players; so the three conjectures form a triple of probability mixtures of action *pairs*. On the other hand, an equilibrium is a triple of mixed actions. Our discussion hinges on the relation between these two kinds of objects.

First, since our real concern is with mixtures of *actions* rather than of action *pairs*, could we not formulate conditions that deal directly with each player's "individual" conjectures—his conjectures about each of the other players—rather than with his overall conjecture? For example, one might hope that it would be sufficient to assume common knowledge of each player's individual conjectures.

Example 41.3 shows that this hope is vain, even when the priors are common and rationality is commonly known. Overall conjectures do play an essential role.

Nevertheless, *common* knowledge of the overall conjectures seems a rather strong assumption. Couldn't we get away with less—say, with mutual knowledge of the overall conjectures, or with mutual knowledge of a high order?

Again, the answer is no. In Example 41.4, there is mutual knowledge of overall conjectures (which may be of an arbitrarily high order), common knowledge of rationality, and common prior, but the individual conjectures do not constitute a Nash equilibrium.

What drives this example is that different players have different individual conjectures about some particular player j, so there isn't even a clear *candidate*

for a Nash equilibrium.[16] This raises the question of whether (sufficiently high order) mutual knowledge of the overall conjectures implies Nash equilibrium of the individual conjectures when the players do happen to agree, in *addition* to assuming (sufficiently high order) mutual knowledge of the overall conjectures. Do we get Nash equilibrium?

Again, the answer is no; this is shown in Example 41.5.

Finally, Example 41.6 shows that the common prior assumption is really needed; it exhibits a situation with common knowledge of the overall conjectures and of rationality, where the individual conjectures agree; but there is no common prior, and the agreed-upon individual conjectures do *not* form a Nash equilibrium.

Summing up, one must consider the overall conjectures, and nothing less than common knowledge of these conjectures, together with common priors, will do.

Except in Example 41.6, the belief systems in this section have common priors, and these are used to describe them. In all the examples, the game being played is (like in section "Illustrations") fixed throughout the belief system, and so is commonly known. Each example has three players, Rowena, Colin, and Matt, who choose the row, column, and matrix (west or east) respectively. As in section "Illustrations," each type is denoted by the same letter as its action, and a subscript is added.

Example 41.3 Here the individual conjectures are commonly known and agreed upon, rationality is commonly known, and there is a common prior, and yet we don't get Nash equilibrium. Consider the game of Fig. 41.7, with theories induced by the common prior in Fig. 41.8.

At each state, Colin and Matt agree on the conjecture $\frac{1}{2}U + \frac{1}{2}D$ about Rowena, and this is commonly known. Similarly, it is commonly known that Rowena and Matt agree on the conjecture $\frac{1}{2}L + \frac{1}{2}R$ about Colin, and Rowena and Colin agree on $\frac{1}{2}W + \frac{1}{2}E$ about Matt. All players are rational at all states, so rationality is common knowledge at all states. But $(\frac{1}{2}U + \frac{1}{2}D, \frac{1}{2}L + \frac{1}{2}R, \frac{1}{2}W + \frac{1}{2}E)$ is not a Nash equilibrium, because if these were independent mixed strategies, Rowena could gain by moving to D.

Fig. 41.7

		L	R
U		1, 1, 1	0, 0, 0
D		1, 0, 0	1, 1, 1

W

		L	R
U		0, 0, 0	1, 1, 1
D		1, 1, 1	0, 0, 0

E

[16] It is *not* what drives Example 41.3; since the individual conjectures are commonly known there, they must agree (Aumann 1976).

Fig. 41.8

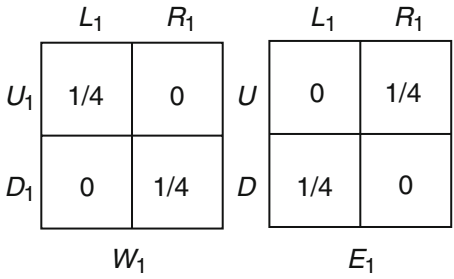
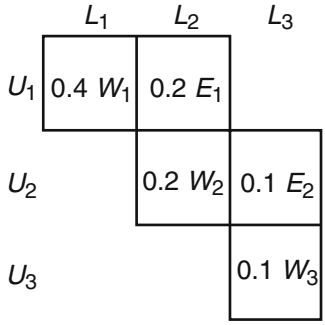


Fig. 41.9



Note that the overall conjectures are not commonly (nor even mutually) known at any state. For example, at (U_1, L_1, W_1), Rowena's conjecture is $(\frac{1}{2}LW + \frac{1}{2}RE)$, but nobody else knows that that is her conjecture.

Example 41.4 Here we have mutual knowledge of the overall conjectures, common knowledge of rationality, and common priors; yet individual conjectures don't agree, so one can't even identify a candidate for a Nash equilibrium.

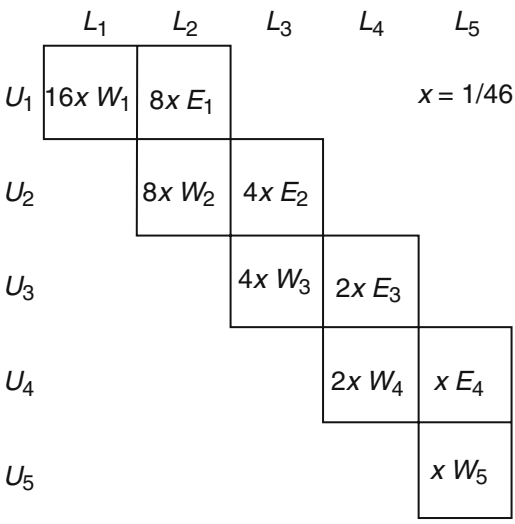
Consider a three-person game in which Rowena and Colin each have just one action—say U for Rowena and L for Colin—and Matt has two actions, W and E. The payoffs are unimportant in this example,[17] since we are only interested in showing that the individual conjectures do not agree. Let the belief system be as in Fig. 41.9. As usual, Rowena's types are depicted by rows, Colin's by columns. Matt's types are indicated in the individual boxes in the diagram; note that in this case, he knows the true state.

Consider the state (U_2, L_2, W_2), or simply W_2 for short. At this state, Rowena's conjecture is $\frac{2}{3}LW + \frac{1}{3}LE$, Colin's is $\frac{1}{2}UW + \frac{1}{2}UE$, and Matt's is UL. Rowena knows Colin's and Matt's conjectures, as they are the same at the only other state (E_2) that she considers possible.[18] Similarly, Colin knows Rowena's and Matt's conjectures, as they are the same at the only other state (E_1) that he considers possible. Matt knows Rowena's and Colin's conjectures, since he knows that the true state is W_2.

[17] They can easily be chosen so that rationality is common knowledge.

[18] I.e., to which she assigns positive probability.

Fig. 41.10



So the conjectures are mutually known. Yet Rowena's conjecture for Matt, $\frac{2}{3}W+\frac{1}{3}E$, is different from Colin's, $\frac{1}{2}W + \frac{1}{2}E$.

The same idea yields examples of this kind with higher order mutual knowledge of the conjectures. See Fig. 41.10. At W_3, there is third order mutual knowledge of the conjectures. By lengthening the staircase[19] and choosing a state in its middle, one can get mutual knowledge of arbitrarily high order that Rowena's conjecture for Matt is $\frac{2}{3}W + \frac{1}{3}E$, while Colin's is $\frac{1}{2}W + \frac{1}{2}E$.

The perspicacious reader will have realized that this example is not intrinsically game-theoretic. It really boils down to a question about "agreeing to disagree": Suppose that two people with the same prior get different information, and that given this information, their posterior probabilities for some event A are mutual knowledge of some order. Are the posterior probabilities then equal? (Here the individuals are Rowena and Colin, and the event A is that Matt chooses W; but actually that's just window dressing.) An example in Aumann (1976) provides a negative answer to the question in the case of first order mutual knowledge. Geanakoplos and Polemarchakis (1982) showed that the answer is negative also for higher order mutual knowledge. The ingenious example presented here is due to John Geanakoplos.[20]

Example 41.5 Here we have mutual knowledge of the overall conjectures, agreement of individual conjectures, common knowledge of rationality, and a common

[19] Alternatively, one can use a single probability system with countably many states, represented by a staircase anchored at the top left and extending infinitely downwards to the right (Jacob's ladder?). By choosing a state sufficiently far from the top, one can get mutual knowledge of any given order.

[20] Private communication. The essential difference between the 1982 example of Geanakoplos and Polemarchakis and the above example of Geanakoplos is that in the former, Rowena's and Colin's

Fig. 41.11

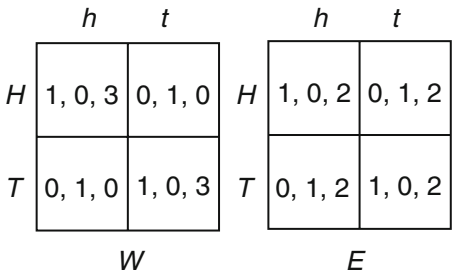
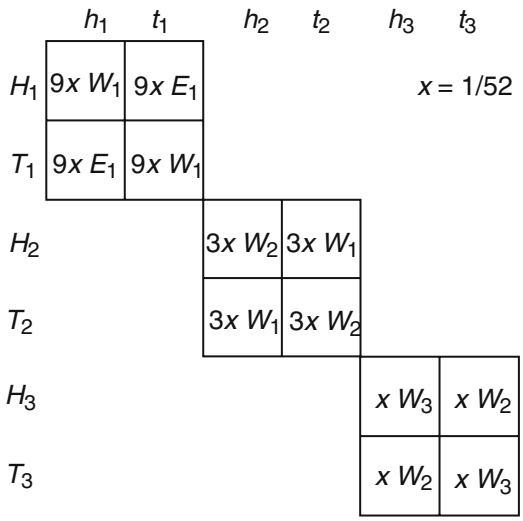


Fig. 41.12



prior, and yet the individual conjectures do not form a Nash equilibrium. Consider the game of Fig. 41.11. For Rowena and Colin, this is simply "matching pennies" (Fig. 41.4); their payoffs are not affected by Matt's choice. So at a Nash equilibrium, they must play $\frac{1}{2}H + \frac{1}{2}T$ and $\frac{1}{2}h + \frac{1}{2}t$ respectively. Thus Matt's expected payoff is 3/2 for W, and 2 for E; so he must play E. Hence $(\frac{1}{2}H + \frac{1}{2}T, \frac{1}{2}h + \frac{1}{2}t, E)$ is the unique Nash equilibrium of this game.

Consider now the theories induced by the common prior in Fig. 41.12. Rowena and Colin know which of the three "boxes" contains the true state, and in fact this is commonly known between the two of them. In each box, Rowena and Colin "play matching pennies optimally"; their conjectures about each other are $\frac{1}{2}H + \frac{1}{2}T$ and $\frac{1}{2}h + \frac{1}{2}t$. Since these conjectures obtain at each state, they are commonly known (among all three players); so it is also commonly known that Rowena and Colin are rational.

probabilities for A approach each other as the other m of mutual knowledge approaches ∞, whereas in the latter, they remain at 2/3 and 1/2 no matter how large m is.

As for Matt, suppose first that he is of type W_1 or W_2. Each of these types intersects two adjacent boxes in Fig. 41.12; it consists of the diagonal states in the left box and the off-diagonal ones in the right box. The diagonal states on the left have equal probability, as do the off-diagonal ones on the right; but on the left, the probability is three times that on the right. So Matt assigns the diagonal states three times the probability of the off-diagonal states; i.e., his conjecture is $\frac{3}{8}Hh + \frac{3}{8}Tt + \frac{1}{8}Th + \frac{1}{8}Ht$. Therefore his expected payoff from choosing W is $\frac{3}{8}3 + \frac{3}{8}3 + \frac{1}{8}0 + \frac{1}{8}0 = 2\frac{1}{4}$, whereas from E it is only 2 (as all his payoffs in the eastern matrix are 2). So W is indeed the optimal action of these types; so they are rational. It may be checked that also E_1 and W_3 are rational. Thus the rationality of all players is commonly known at all states.

Consider now the state $s := (H_2, h_2, W_2)$ (the top left state in the middle box). Rowena and Colin know at s that they are in the middle box, so they know that Matt's type is W_1 or W_2. We have just seen that these two types have the same conjecture, so it follows that Matt's conjecture is mutually known at s. Also Rowena's and Colin's conjectures are mutually known at s (Rowena's is $\frac{1}{2}hW + \frac{1}{2}tW$, Colin's is $\frac{1}{2}HW + \frac{1}{2}TW$).

Finally, the individual conjectures derived from Matt's overall conjecture $\frac{3}{8}Hh + \frac{3}{8}Tt + \frac{1}{8}Th + \frac{1}{8}Ht$ are $\frac{1}{2}H + \frac{1}{2}T$ for Rowena and $\frac{1}{2}h + \frac{1}{2}t$ for Colin. These are the same as Rowena's and Colin's conjectures for each other. Also, since Matt plays W throughout the middle box, both Rowena and Colin conjecture W for Colin there. Thus throughout the middle box, individual conjectures are agreed upon.

To sum up: There is a common prior; at all states, the game is commonly known and all players are commonly known to be rational. At the top left state in the middle box, the overall conjectures of all players are mutually known, and the individual conjectures are agreed: $\sigma_R = \frac{1}{2}H + \frac{1}{2}T$, $\sigma_C = \frac{1}{2}h + \frac{1}{2}t$, $\sigma_M = W$. But $(\sigma_R, \sigma_C, \sigma_M)$ is not a Nash equilibrium.

To understand the construction, note that in the east matrix, Matt gets 2 no matter what the others do; in the west matrix, he gets 3 if the others play on the (main) diagonal, 0 if they play off. Thus he is motivated to play W when his conjecture assigns a high probability to their playing on the diagonal. The common prior in Fig. 41.12 generates such a conjecture that is mutually known, in spite of the overall probability of the diagonal being only 1/2. The reason for using the diagonal[21] is to avoid correlation with Rowena's or Colin's information, so as to assure that Matt's conjectures about them will agree with their conjectures about each other.

As in Example 41.4, one can construct a similar example for which the mutual knowledge of the conjectures is of an arbitrarily high order, simply by using more boxes; the result follows as before.

[21] Rather than the upper row, say.

Fig. 41.13

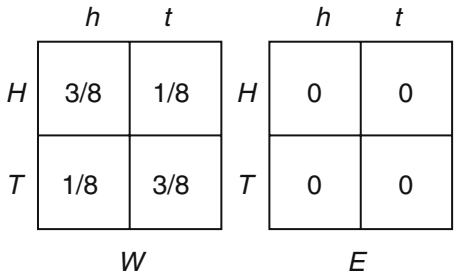
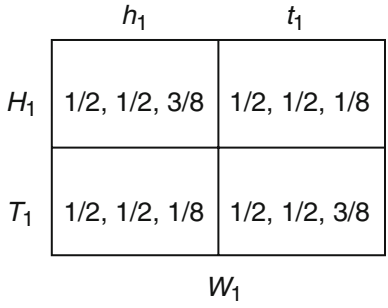


Fig. 41.14



Agreement between individual conjectures may be viewed as a kind of "consistency" between the overall conjectures φ^i. A stronger consistency condition,[22] suggested by Kenneth Arrow, is that there be a single probability distribution ϕ on $A = \mathrm{x}_{i=1}^{n} A_i$ whose marginals on the spaces A^{-i} are the φ^i. One may ask whether with this stronger condition, one can replace common by mutual knowledge in Theorem 41.3. The answer is still no, even with mutual knowledge of arbitrarily high order. The above example satisfies Arrow's condition, with $\phi = \frac{3}{8}HhW + \frac{3}{8}TtW + \frac{1}{8}ThW + \frac{1}{8}HtW$ (Fig. 41.13).

Example 41.6 Here we show that one cannot dispense with common priors in Theorem 41.3. Consider again the game of Fig. 41.11, with the theories depicted in Fig. 41.14 (presented in the style of Figs. 41.2 and 41.5; note that Matt has no type[23] whose action is E). At each state there is common knowledge of rationality, of overall conjectures (which are the same as in the previous example), and of the game. As before, Arrow's condition is satisfied, and it follows that the individual conjectures are in agreement. And as before, the individual conjectures $(\frac{1}{2}H + \frac{1}{2}T, \frac{1}{2}h + \frac{1}{2}t)$ do not constitute a Nash equilibrium.

[22] Though similar in form, this condition neither implies nor is implied by common priors. We saw in Example 41.4 that common priors do not even imply agreement between individual forecasts; a fortiori, they do not imply Arrow's condition. In the opposite direction, Example 41.6 satisfies Arrow's condition, but has no common prior.

[23] If one wishes, one can introduce a type E_1 of Matt to which Rowena's and Colin's types ascribe probability 0, and whose theory is, say, $\frac{1}{4}Hh + \frac{1}{4}Tt + \frac{1}{4}Th + \frac{1}{4}Ht$.

Fig. 41.15

	h_1	t_1
H_1	1/6	1/6
T_1	1/3	1/3

Fig. 41.16

Game T

	S
U	1, 1
D	0, 0

Game B

	S
U	0, 1
D	1, 0

Additional Counterexamples

In the previous section we saw that the assumption of a common prior and of common knowledge of the conjectures are essential in Theorem 41.3. This section explores the assumption of mutual knowledge of rationality and of the game being played (in both Theorems 41.2 and 41.3), and shows that they, too, cannot be substantially weakened.

Example 41.7 Here we show that in Theorems 41.2 and 41.3, mutual knowledge of rationality cannot be replaced by the simple fact of rationality (as in Theorem 41.1). Consider again the game of "matching pennies" (Fig. 41.4), this time with theories induced by the common prior in Fig. 41.15.

At the state (H_1, h_1), Colin's and Rowena's conjectures are commonly known to be $\frac{1}{3}H + \frac{2}{3}T$ and $\frac{1}{2}h + \frac{1}{2}t$ respectively, and both are in fact rational (indeed Rowena's rationality is commonly known); but $(\frac{1}{3}H + \frac{2}{3}T, \frac{1}{2}h + \frac{1}{2}t)$ is not a Nash equilibrium, since the only equilibrium of "matching pennies" is $(\frac{1}{2}H + \frac{1}{2}T, \frac{1}{2}h + \frac{1}{2}t)$. Note that at the state (H_1, h_1), Rowena does not know that Colin is rational.

Example 41.8 Here we show that in Theorems 41.2 and 41.3, knowing one's own payoff function does not suffice; one needs mutual knowledge of all the payoff functions. Consider a two-person belief system where one of two games, T (top) or B (bottom), is being played (Fig. 41.16).

Fig. 41.17

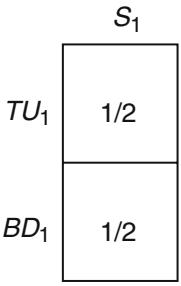
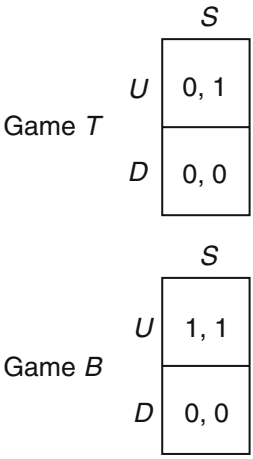


Fig. 41.18



The theories are given by the common prior in Fig. 41.17. Thus Rowena knows Colin's payoff function, but Colin does not know Rowena's. Rowena's type TU_1 has the *Top* payoff function and plays *Up*, whereas BD_1 has the *Bottom* payoff function and plays *Down*. Colin has just a single type, S_1. At both states, both players are rational: Rowena, who knows the game, always plays an action that is strictly dominant in the true game; Colin has no choice, so what he does is rational. So there is common knowledge of rationality. At both states, Colin's conjecture about Rowena is $\frac{1}{2}U + \frac{1}{2}D$; Rowena's conjecture about Colin is S. But $(\frac{1}{2}U + \frac{1}{2}D, S)$ is not a Nash equilibrium in either of the games; Rowena prefers U in the top game, D in the bottom game.

Example 41.9 Here we show that the hypotheses of Theorem 41.3 do not imply that rationality is commonly known (as is the case when the game g is commonly known—see Proposition 41.A1). Consider a two-person belief system where one of the two games of Fig. 41.18 is being played. The theories are given by the common prior in Fig. 41.19. Of Rowena's three types, TU_1 and TD_1 are rational, whereas BD_1 (who plays *Down* in Game B) is irrational. Colin's two types, S_1 and S_2, differ only in their theories, and both are rational. The conjectures $\frac{1}{2}U + \frac{1}{2}D$ and S are common knowledge at all states. At the state (TU_1, S_1), it is mutually known that T

Fig. 41.19

	S_1	S_2
TU_1	1/4	1/4
TD_1	1/4	0
BD_1	0	1/4

is the game being played, and that both players are rational; but neither rationality nor the game being played are commonly known.

General (Infinite) Belief Systems

For a general definition of a belief system, we specify that the type spaces S_i be measurable spaces. As before, a *theory* is a probability measure on $S^{-i} = \times_{j \neq i} S_j$, which is now endowed with the standard product structure.[24] The state space $S = \times_j S_j$, too, is endowed with the product structure. An *event* is now a measurable subset of S. The "action functions" $\mathbf{a}_i$ ((3)) are assumed measurable; so are the payoff functions $\mathbf{g}_i$ ((4)), as functions of s_i, for each action n-tuple a separately. Also the "theory functions" ((2)) are assumed measurable, in the sense that for each event E and player i, the probability $p(E; s_i)$ is measurable as a function of the type s_i. It follows that also the conjectures $\boldsymbol{\varphi}^i$ are measurable functions of s_i.

With these definitions, the statements of the results make sense, and the proofs remain correct, without any change.

Discussion

(a) Belief Systems

An interactive belief system is not a prescriptive model; it does not suggest actions to the players. Rather, it is a formal framework—a language—for *talking* about actions, payoffs, and beliefs. For example, it enables us to say whether a given player is behaving rationally at a given state, whether this is known to another player, and

[24] The σ-field of measurable sets is the smallest σ-field containing all the "rectangles" $\times_{j \neq i} T_j$, where T_j is measurable in S_j.

so on. But it does not prescribe or even suggest rationality; the players do whatever they do. Like the disk operating system of a personal computer, the belief system simply organizes things, so that we can coherently discuss what they do.

Though entirely apt, use of the term "state of the world" to include the actions of the players has perhaps caused confusion. In Savage (1954), the decision maker cannot affect the state; he can only react to it. While convenient in Savage's one-person context, this is not appropriate in the interactive, many-person world under study here. Since each player must take into account the actions of the others, the actions should be included in the description of the state. Also the plain, everyday meaning of the term "state of the world" includes one's actions: Our world is shaped by what we do.

It has been objected that prescribing what a player must do at a state takes away his freedom. This is nonsensical; the player may do what he wants. It is simply that whatever he does is part of the description of the state. If he wishes to do something else, he is heartily welcome to do it, but he thereby changes the state.

Historically, belief systems were introduced by John Harsanyi (1967–1968), to enable a coherent formulation of games in which the players need not know each other's payoff functions. To analyze such games, it is not enough to specify each player's beliefs about (i.e., probability distributions on) the payoff functions of the others; one must also specify the beliefs of the players about the beliefs of the players about the payoff functions, the beliefs of the players about *these* beliefs, and so on ad infinitum. This complicated infinite regress seemed to make useful analysis very difficult.

Harsanyi's ingenious solution was to think of each player as being one of several possible "types," where a type determines both a player's own payoff function and a belief about the types of the others. The belief of a player about the types of the others induces a belief about their payoff functions; it also induces a belief about their beliefs about the types, and so a belief about the beliefs about the payoff functions. The reasoning continues indefinitely. Thus from an *I-game* ("I" for incomplete information), as Harsanyi calls his type-model, one can read off the entire infinite regress of beliefs.

Belief systems as defined in section "Interactive belief systems" are formally just like Harsanyi's I-games, except that in belief systems, a player's type determines his action as well as his payoff function and his belief about other players' types.[25] As above, it follows that the player's type determines his entire *belief hierarchy*—i.e., the entire infinite regress of his beliefs about actions, beliefs about beliefs about actions, and so on—in addition to the infinite regress of beliefs about payoff functions, and how these two kinds of beliefs affect each other.

Traditionally, payoff functions have been treated as exogenous, actions as endogenous. It was thought that unlike payoff functions, actions should be "predicted" by the theory. Belief systems wipe out this distinction; they treat uncertainty

[25]For related ideas, see Armbuster and Boege (1979), Boege and Eisele (1979), and Tan and Werlang (1988).

about actions just like uncertainty about payoff functions. Indeed, in this paper the focus is on actions;[26] uncertainty about payoff functions was included as an afterthought, because we realized that more comprehensive results can be obtained at almost no cost in the complexity of the proofs.

(b) Common Knowledge of the Belief System

Is the belief system itself common knowledge among players? If so, how does it get to be common knowledge? If not, how do we take into account the players' uncertainty about it?

A related question is whether the belief system is exogenous, like a game or a market model. If not, where does it come from?

The key to these issues was provided[27] in a fundamental paper of Mertens and Zamir (1985). They treat the Harsanyi case, in which the "underlying variables"[28] are the payoff functions only (see (a) above); but the result applies without change to our situation, where actions as well as payoff functions are underlying variables.

At (a) above, we explained how each type in a belief system determines a belief hierarchy. Mertens and Zamir reverse this procedure: They start with the belief hierarchies, and construct belief systems from them. Specifically, they define the *universal belief space* as a belief system in which the type space of each player is simply the set of *all* his belief hierarchies that satisfy certain minimal consistency conditions. Thus the universal belief space is not exogenously given, like a game or market; rather, it is an analytic tool, like the payoff matrix of a game originally given in extensive form. It follows that the universal belief space may be considered common knowledge.

Though the universal belief space is infinite, it is the disjoint union of infinitely many "common knowledge components,"[29] many of which are finite. Mertens and Zamir call any union of such components a *subspace* of the universal belief space. It follows that when the belief system is a subspace, then *the belief system itself is common knowledge.*

It may be shown that *any* belief system B for $A_1 \times \cdots\cdots \times A_n$—including, of course, a finite one—is "isomorphic" to a subspace of the universal belief space.[30] From all this we conclude that the belief system itself may *always* be considered common knowledge.

---

[26] As is apparent from the examples in sections "Illustrations," "The main counterexamples," and "Additional counterexamples," in most of which the game g is common knowledge.

[27] See also Armbuster and Boege (1979) and Boege and Eisele (1979).

[28] The variables about which beliefs—of all orders—are held.

[29] At each state s of such a component S, the identity of that component is commonly known (i.e., it is commonly known at s that the true state is in S, though it need not be commonly known that the true state is s).

[30] This is because each type in B determines a belief hierarchy (see (a)). The set of n-tuples of all these belief hierarchies is the subspace of the universal belief space that is isomorphic to B.

(c) Knowledge and Belief

In this paper, "know" means "ascribe probability 1 to." This is sometimes called "believe," while "know" is reserved for absolute certainty, with no possibility at all for error. In the formalism of belief systems[31] (section "Interactive belief systems"), absolute certainty has little concrete meaning; a player can be absolutely certain only on his own action, his own payoff function, and his own theory, not of anything pertaining to anybody else.

Choosing between the terms "know" and "believe" caused us many sleepless nights. In the end, we decided on "know" because it enables simpler, less convoluted language, and because we were glad to de-emphasize the relatively minor conceptual difference between probability 1 and absolute certainty.[32]

Note that since our conditions are sufficient, our results are stronger with probability 1 than with absolute certainty. If probability 1 knowledge of certain events implies that σ is a Nash equilibrium, then a fortiori, so does absolute certainty of those events.

(d) Structure of the Game

In section "Introduction," we identified the "structure of the game" with an n-tuple g of payoff functions. As pointed out by John Harsanyi (1967–1968), this involves no conceptual loss of generality, as one can always pick the action sets A_j sufficiently large to justify this.

(e) Alternative Formalisms

Instead of using belief systems to formulate and prove our results and examples, one may use knowledge partitions (e.g., Aumann 1976). The advantage of the partition formalism is that with it one can represent absolute certainty, not just probability 1 knowledge (see (c)). Also some of the proofs may become marginally simpler.[33] On the other hand, the partition formalism—especially when combined with probabilities—is itself more complex than that of belief systems, and it is desirable to use as simple, transparent, and unified an approach as possible.

[31] Unlike that of partitions (see (e) below).

[32] Another reason is that "belief" often refers to a general probability distribution, which does not go well with using "know" to mean "ascribe probability 1 to."

[33] This is natural, as the results are slightly weaker (see (c)).

(f) Independent Conjectures

Theorem 41.3 deduces independence of the individual conjectures from common knowledge of the overall conjectures. An alternative approach is embodied in the following:

Remark 41.1 *Let σ be an n-tuple of mixed strategies. Suppose that at some state, it is mutually known that the players are rational, that the game g is being played, that the conjecture of each player i about each other player j is σ_j, and that it is independent of i's conjecture about all other players. Then σ is a Nash equilibrium in g.*

The proof is as in the last part of Theorem 41.3's proof, after (41.3) has been established.

Here we assume mutual rather than common knowledge of conjectures and do not assume common priors. On the other hand, we assume outright that the individual conjectures are agreed upon, and that each player's conjectures about the others are independent. Because of the strength of these assumptions, the result is of limited interest.

Mutual independence of the conjectures of a player *i* about different players *j* seems a particularly untenable assumption. It may well be that Colin and Matt "choose their actions independently" in the sense of not consulting each other or even communicating, but that by no means implies that Rowena's conjecture about Colin is stochastically independent of her conjecture about Matt. It is possible to have considerable stochastic dependence even when there is no possibility of communication or correlation between Colin and Matt—even when they are placed in separate rooms and then informed of the game they are to play. The probability is in Rowena's head, it is a question of *her* beliefs, *her* ignorance or knowledge, it is not directly determined by some kind of physical process such as "independently" choosing or being in separate rooms or anything like that. Suppose, for example, that Colin and Matt could each act either "boldly" or "carefully." Even if Rowena knows "nothing at all" about Colin and Matt (whatever that might mean; perhaps she believes they came from unspecified different directions in outer space), her probability that Matt acts boldly given that Colin does might well be higher than her unconditional probability. After all, it is a question of *her* outlook on life, how *she* thinks people react to things; if she is at all sophisticated, she might easily say to herself, "Well, I'm not sure of human nature, I don't know what people do in this kind of situation; but if Colin plays boldly, that tells me something about people in general, and it raises my probability that Matt will also play boldly." In fact, it is quite unlikely that the conjectures would be stochastically independent, one would have great difficulty in constructing a set of circumstances that would justify such a conclusion.

When teaching probability, we are at pains to point out that a sequence of coin tosses is *not* in general represented by an i.i.d. sequence of random variables. This would be the case only if the parameter of the coin were known with certainty before

the first toss, a highly unlikely proposition. And this in spite of the fact that the tosses themselves are of course "physically" independent (whatever that may mean; we're not sure it means anything). The relation between "physical" and stochastic independence is murky at best.

Independence of individual conjectures is an appropriate assumption when one thinks of mixed strategies in the old way, in terms of explicit randomizations. With that kind of interpretation, "acting independently" is of course closely associated with stochastic independence. But not when the probabilities represent other players' ignorance.

To be sure, common knowledge of the conjectures is also a rather strong assumption. Moreover, we do *conclude* that the individual conjectures are independent. But there is a difference between an assumption that is merely strong and one that is groundless. More to the point, there is a qualitative difference between assuming common knowledge of the conjectures and assuming independence. Common knowledge of the conjectures describes what might almost be called a "physical" state of affairs. It might, for example, be the outgrowth of experience or learning, like common knowledge of a language. Independence, on the other hand, is in the mind of the decision maker. It isn't reasonable to make such an assumption when one can't describe some clear set of circumstances under which it would obtain—and that is very difficult to do. By contrast, common knowledge of the conjectures may itself be considered a "clear set of circumstances."

(g) Epistemic "Equivalence" Conditions

This paper is devoted to sufficient epistemic conditions for Nash equilibrium. Another kind of epistemic result is illustrated by the following:

Remark 41.2 *Let $\mathbf{a}$ be an n-tuple of actions. Suppose that at some state s, it is commonly known that $\mathbf{g} = g$ and $\mathbf{a} = a$. Then rationality is commonly known at s iff a is a Nash equilibrium in g.*

For another result of this kind, which relates to Theorem 41.2 as the above relates to Theorem 41.1, see Brandenburger and Dekel (1989). These results provide sufficient conditions for the *equivalence* between Nash equilibrium and common knowledge of rationality, rather than directly for Nash equilibrium.

(h) Global Epistemic Conditions

The epistemic conditions discussed in this paper are *local*; they say when the players' actions or conjectures at a specific state constitute an equilibrium. By contrast, one can treat *global* conditions, which refer to the belief system as a whole.

These are usually quite different from the local conditions treated here. For example, the condition for correlated equilibrium in Aumann (1987) is global.

A global epistemic condition for Nash equilibrium[34] is as follows:

Remark 41.3 *Suppose that the game g is fixed,*[35] *that there is a common prior P, and that all players are rational at all states. Then the distribution Q of action n-tuples is a Nash equilibrium in g if and only if for each player i, the expectation of i's conjecture given one of his actions a_i is the same for all a_i that are assigned positive probability by Q.*

Roughly speaking, the condition says that if the players use only information that is relevant to their choice of an action, then their conjectures are always the same, independently of their information.

(i) Irrationality, Probability Assessments, and Payoff Functions

There is a difficulty when explicitly allowing for irrationality, as we do here. The problem rears its ugly head on several levels. Most fundamentally, a player is by definition rational if his choice maximizes his utility function. Again by definition, the utility function represents his preferences, and yet again by definition, the player prefers x to y iff he chooses x when given the choice between x and y. So "rationality" appears tautologous: utility is always maximized; it is *defined* as a function that the choice maximizes.

One may attempt to resolve this by noting that whereas rationality requires the choice x to maximize the utility function, that does not mean that any x maximizes some utility function. Indeed there may not be any utility function at all. Rationality must be understood in terms of a richer structure, one involving preferences (choices between pairs or from other subsets of the available alternatives). If these preferences do not satisfy basic requirements (such as transitivity), then we may speak of irrationality.

But the difficulty reappears at a different level. In a belief system, all players, rational or not, have theories—i.e., probability assessments over the types of others, and so over their actions and so on. In the foundations of decision theory, e.g., à la Savage, probabilities and utilities are derived from axioms that apply to rational players only. How, then, can an irrational player entertain probabilities—indeed, how can he have a payoff function, which is presumably derived from his utility function?

There are several ways to address this problem. One is to change the definition of a belief system, by specifying for each type whether or not it is "rational," and then specifying theories and payoff functions only for types that are specified

[34] Stated, but not proved, in Aumann (1987).

[35] Constant throughout the belief system.

as rational.[36] This would rescue Theorems 41.1 and 41.2, since in those results, probability assessments and payoff functions of irrational types play no role.

It would, however, not rescue Theorem 41.3, since there one needs common knowledge of the conjectures, but only mutual knowledge of rationality. Thus it appears that irrational players must entertain conjectures; for this they should be rational. To resolve this problem, one may distinguish between *subjective* and *objective* theories—and so also between subjective and objective knowledge, conjectures, and rationality. Thus, think of the common prior in Theorem 41.3 as referring to the "objective" assessment of a fixed outside observer Otto. Call type s_i *objectively* rational if its choice maximizes its expected payoff when calculated according to the objective conjecture (the one that Otto would hold if given s_i's information), and according to its own payoff function. This differs from *subjective* rationality of s_i (the notion of rationality hitherto used), which requires that s_i's choice be maximal according to s_i's own conjecture, not Otto's. Similarly, say that a type s_i knows some event E *objectively* if given s_i's information, Otto would ascribe probability 1 to E. Theorem 41.3 can then be understood as asking for mutual objective knowledge of the players' payoff functions and of objective rationality, and for common objective knowledge of objective conjectures. These assumptions do not demand that irrational types entertain conjectures, but they may. If they do, they are now to be thought of as subjectively—but not necessarily objectively—rational.

Needless to say, our results remain true—and almost as interesting—when rationality *is* commonly known. Thus readers who are dissatisfied with the above interpretations may simply assume common knowledge of rationality.

(j) Knowledge of Equilibrium

The conclusions of our theorems state that a specified (mixed) strategy n-tuple σ is an equilibrium; they do not state that the players know it to be an equilibrium, or that this is commonly known. In the case of Theorems 41.2 and 41.3, though, it is in fact mutual knowledge of order 1—but not necessarily of any higher order—that σ is a Nash equilibrium. In the case of Theorem 41.1, it need not even be mutual knowledge of order 1 that σ is a Nash equilibrium; but this does follow if, in addition to the stated assumptions, one assumes mutual knowledge of the payoff functions.

[36] It would be desirable to see whether and how one can derive this kind of modified belief system from a Mertens–Zamir-type construction.

(k) Conclusions

Where does all this leave us? Are the "foundations" of Nash equilibrium more secure or less secure than previously imagined? Do our results strengthen or weaken the case for Nash equilibrium?

First, in assessing the validity of a game-theoretic solution concept, one should not place undue emphasis on its a priori rationale. At least as important is the question of how successful the concept is in providing insight into applications, how tractable it is, and, relatedly, even the extent of its aesthetic appeal. On all these counts, Nash equilibrium has proved its worth.

This said, the present results do indicate that the a priori case for Nash equilibrium is a little stronger than conventional wisdom has granted. Common knowledge turns out to play a more limited role—at least in games with two players—than previously thought. The reader may object that even mutual knowledge of, say, payoff functions is implausible; but indisputably, common knowledge of payoff functions is more so. It is true that our epistemic conditions for Nash equilibrium in games with more than two players involve common knowledge (of conjectures); indeed, it was surprising to discover that the conditions for equilibrium in the n-person case are stronger than in the two-person case. Perhaps Nash equilibrium rests on firmer foundations in two-person than in n-person games.

It should be remembered that the conditions for Nash equilibrium described here are sufficient, but not necessary. Perhaps there are other ways of looking at Nash equilibrium epistemically; if so, their nature is as yet unclear.

There also are non-epistemic ways of looking at Nash equilibrium, such as the evolutionary approach (e.g., Maynard Smith 1982). Related to this is the idea that a Nash equilibrium represents a societal norm. In the end, these viewpoints will perhaps provide a more compelling basis for Nash equilibrium than those involving games played by consciously maximizing, rational players.

Finally, the apparatus of this paper—belief systems, conjectures, knowledge, mutual knowledge, common knowledge, and the like—has an appeal that extends beyond our immediate purpose of providing epistemic conditions for Nash equilibrium. The apparatus offers a way of analyzing strategic situations that corresponds nicely to the concerns that we have all experienced in practice—what is the other person thinking, what is his true motivation, does he see the world as I do, and so on.

Appendix: Extensions and Converses

We start with some remarks on Theorem 41.3 and its proof. First, the conclusions of Theorem 41.3 continue to hold under the slightly weaker assumption that the common prior assigns positive probability to $\boldsymbol{\varphi} = \varphi$ being commonly known, and there is a state at which $\boldsymbol{\varphi} = \varphi$ is commonly known and $\mathbf{g} = g$ and the rationality of the players are mutually known.

Second, note that the rationality assumption is not used until the end of Theorem 41.3's proof, after (41.3) is established. Thus if we assume only that there is a common prior that assigns positive probability to the conjectures φ^i being commonly known, we may conclude that all players i have the same conjecture σ_j for other players j, and that each φ^i is the product of the σ_j with $j \neq i$; that is, the $n-1$ conjectures of each player about the other players are independent.

Third, if in Theorem 41.3 we assume that the game being played is commonly (not just mutually) known, then we can conclude that also the rationality of the players is commonly known.[37] That is, we have

Proposition 41.A1 *Suppose that at some state s, the game g and the conjectures φ^i are commonly known and rationality is mutually known. Then at s, rationality is commonly known. (Note that common priors are not assumed here.)*

Proof Set $G := [g]$, $F := [\varphi]$, $R_j := [j \text{ is rational}]$, and $R := [\text{all players are rational}] = R_1 \cap \cdots \cap R_n$. In these terms, the proposition says that $CK(G \cap F) \cap K^1 R \subset CKR$. We assert that it is sufficient for this to prove

$$K^2(G \cap F) \cap K^1 R \subset K^2 R. \tag{41.A1}$$

Indeed, if we have (A1), an inductive argument using Lemma 41.5 and that $E \subset E'$ implies $K^1(E) \subset K^1(E')$ (which follows from Lemma 41.2) yields $K^m(G \cap F) \cap K^1 R \subset K^m R$ for any m; so taking intersections, $CK(G \cap F) \cap K^1 R \subset CKR$ follows.

Let j be a player, B_j the set of actions a_j of j to which the conjecture φ^i of some other player i assigns positive probability. Let $E_j := [\mathbf{a}_j \in B_j]$ (the event that the action chosen by j is in B_j). Since the game g and the conjectures φ^i are commonly known at s, they are a fortiori mutually known there; so by Lemma 41.6, each action in B_j maximizes g_j against φ^j. Hence $E_j \cap G \cap F \subset R_j$. At each state in F, each player other than j knows that j's action is in B_j; that is, $F \subset \cap_{i \neq j} K_i E_j$. So $G \cap F \subset (\cap_{i \neq j} K_i E_j) \cap (G \cap F)$. So Lemmas 41.2 and 41.5 yield

$$\begin{aligned} K^2(G \cap F) &\subset K^2\left(\cap_{i \neq j} K_i E_j\right) \cap K^2(G \cap F) \subset K^1\left(\cap_{i \neq j} K_i E_j\right) \cap K^2(G \cap F) \subset \\ K^1\left(\cap_{i \neq j} K_i E_j\right) &\cap K^1\left(\cap_{i \neq j} K_i (G \cap F)\right) = K^1\left(\cap_{i \neq j} K_i \left(E_j \cap G \cap F\right)\right) \subset \\ &\times K^1\left(\cap_{i \neq j} K_i R_j\right). \end{aligned}$$

Hence using $R \subset R_j$ and $R_j = K_j R_j$ (Corollary 41.2), we obtain

$$\begin{aligned} K^2(G \cap F) \cap K^1 R &\subset K^1\left(\cap_{i \neq j} K_i R_j\right) \cap K^1 R \subset K^1\left(\cap_{i \neq j} K_i R_j\right) \cap K^1 R_j = \\ K^1\left(\cap_{i \neq j} K_i R_j\right) &\cap K^1 K_j R_j = K^2 R_j. \end{aligned}$$

Since this holds for all j, Lemma 41.2 yields[38] (A.2). ■

[37] This observation, for which we are indebted to Ben Polak, is of particular interest because in many applied contexts there is only one game under consideration, so it is of necessity commonly known.

[38] The proof would be simpler with a formalism in which known events are true ($K_j E \subset E$). See section "Alternative formalisms."

Our fourth remark is that in both Theorems 41.2 and 41.3, mutual knowledge of rationality may be replaced by the assumption that each player knows the others to be rational; in fact, all players may themselves be irrational at the state in question. (Recall that "know" means "ascribe probability 1"; thus a player may be irrational even though another player knows that he is rational.)

We come next to the matter of converses to our theorems. We have already mentioned (at the end of section "Description of the results") that the conditions are not necessary, in the sense that it is quite possible to have a Nash equilibrium even when they are not fulfilled. In Theorem 41.1, the action n-tuple $\mathbf{a}(s)$ at a state s may well be a Nash equilibrium even when $\mathbf{a}(s)$ is not mutually known, whether or not the players are rational. (But if the actions are mutually known at s and $\mathbf{a}(s)$ is a Nash equilibrium, then the players *are* rational at s; cf. Remark 41.2.) In Theorem 41.2, the conjectures at a state s in a two-person game may constitute a Nash equilibrium even when, at s, they are not mutually known and/or rationality is not mutually known. Similarly for Theorem 41.3.

Nevertheless, there is a sense in which the converses hold: Given a Nash equilibrium in a game g, one can construct a belief system in which the conditions are fulfilled. For Theorem 41.1, this is immediate: Choose a belief system where each player i has just one type, whose action is i's component of the equilibrium and whose payoff function is g_i. For Theorems 41.2 and 41.3, we may suppose that as in the traditional interpretation of mixed strategies, each player chooses an action by an independent conscious randomization according to his component σ_i of the given equilibrium σ. The types of each player correspond to the different possible outcomes of the randomization; each type chooses a different action. All types of player i have the same theory, namely, the product of the mixed strategies of the other $n-1$ players appearing in σ, and the same payoff function, namely g_i. It may then be verified that the conditions of Theorems 41.2 and 41.3 are met.

These "converses" show that the sufficient conditions for Nash equilibrium in our theorems are not too strong, in the sense that they do not imply more than Nash equilibrium; every Nash equilibrium is attainable with these conditions. Another sense in which they are not too strong—that the conditions cannot be dispensed with or even appreciably weakened—was discussed in sections "The main counterexamples" and "Additional counterexamples."

References

Armbruster, W., & Boege, W. (1979). Bayesian game theory. In O. Moeschlin & D. Pallaschke (Eds.), *Game theory and related topics*. Amsterdam: North-Holland.

Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics, 4*, 1236–1239.

Aumann, R. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica, 55*, 1–18.

Boege, W., & Eisele, T. (1979). On solutions of Bayesian games. *International Journal of Game Theory, 8*, 193–215.

Brandenburger, A., & Dekel, E. (1989). The role of common knowledge assumptions in game theory. In F. Hahn (Ed.), *The economics of missing markets, information, and games*. Oxford: Oxford University Press.

Geanakoplos, J., & Polemarchakis, H. (1982). We can't disagree forever. *Journal of Economic Theory, 28*, 192–200.

Harsanyi, J. (1967–1968). Games of incomplete information played by 'Bayesian' players, I-III. *Management Science*, *14*, 159–182, 320–334, 486–502.

Harsanyi, J. (1973). Games with randomly disturbed payoffs: A new rationale for mixed strategy equilibrium points. *International Journal of Game Theory, 2*, 1–23.

Kohlberg, E., & Mertens, J.-F. (1986). On the strategic stability of equilibria. *Econometrica, 54*, 1003–1037.

Kreps, D., & Wilson, R. (1982). Sequential equilibria. *Econometrica, 50*, 863–894.

Lewis, D. (1969). *Conventions: A philosophical study*. Cambridge: Harvard University Press.

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge: Cambridge University Press.

Mertens, J.-F., & Zamir, S. (1985). Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory, 14*, 1–29.

Myerson, R. (1978). Refinements of the Nash equilibrium concept. *International Journal of Game Theory, 7*, 73–80.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics, 54*, 286–295.

Savage, L. (1954). *The foundations of statistics*. New York: Wiley.

Selten, R. (1965). Spieltheoretische Behandlung eines Oligopolmodels mit Nachfragetragheit. *Zietschrift fur die gesante Staatswissenschaft, 121*, 301–324.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory, 4*, 25–55.

Tan, T., & Werlang, S. (1988). The Bayesian foundations of solution concepts of games. *Journal of Economic Theory, 45*, 370–391.

Chapter 42
Knowledge, Belief and Counterfactual Reasoning in Games

Robert Stalnaker

Introduction

Deliberation about what to do in any context requires reasoning about what will or would happen in various alternative situations, including situations that the agent knows will never in fact be realized. In contexts that involve two or more agents who have to take account of each others' deliberation, the counterfactual reasoning may become quite complex. When I deliberate, I have to consider not only what the causal effects would be of alternative choices that I might make, but also what other agents might believe about the potential effects of my choices, and how their alternative possible actions might affect my beliefs. Counterfactual possibilities are implicit in the models that game theorists and decision theorists have developed – in the alternative branches in the trees that model extensive form games and the different cells of the matrices of strategic form representations – but much of the reasoning about those possibilities remains in the informal commentary on and motivation for the models developed. Puzzlement is sometimes expressed by game theorists about the relevance of what happens in a game 'off the equilibrium path': of what would happen if what is (according to the theory) both true and known by the players to be true were instead false. My aim in this paper is to make some suggestions for clarifying some of the concepts involved in counterfactual reasoning in strategic contexts, both the reasoning of the rational agents being modeled, and the reasoning of the theorist who is doing the modeling, and to bring together some ideas and technical tools developed by philosophers and logicians that I think might be relevant to the analysis of strategic reasoning, and more generally to the conceptual foundations of game theory.

R. Stalnaker (✉)
Department of Linguistics and Philosophy, MIT, Cambridge, MA, USA
e-mail: stal@mit.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_42

There are two different kinds of counterfactual possibilities – causal and epistemic possibilities – that need to be distinguished. They play different but interacting roles in a rational agent's reasoning about what he and others will and should do, and I think equivocation between them is responsible for some of the puzzlement about counterfactual reasoning. In deliberation, I reason both about how the world might have been different if I or others did different things than we are going to do, and also about how my beliefs, or others' beliefs, might change if I or they learned things that we expect not to learn. To take an often cited example from the philosophical literature to illustrate the contrast between these two kinds of counterfactual suppositions, compare: *if Shakespeare didn't write Hamlet, someone else did,* with *if Shakespeare hadn't written Hamlet, someone else would have*.[1] The first expresses a quite reasonable disposition to hold onto the belief that someone wrote Hamlet should one receive the unexpected information that Shakespeare did not; the second expresses a causal belief, a belief about objective dependencies, that would be reasonable only if one held a bizarre theory according to which authors are incidental instruments in the production of works that are destined to be written. The content of what is supposed in the antecedents of these contrasting conditionals is the same, and both suppositions are or may be counterfactual in the sense that the person entertaining them believes with probability one that what is being supposed is false. But it is clear that the way it is being supposed is quite different in the two cases.

This contrast is obviously relevant to strategic reasoning. Beliefs about what it is rational to do depend on causal beliefs, including beliefs about what the causal consequences would be of actions that are alternatives to the one I am going to choose. But what is rational depends on what is believed, and I also reason about the way my beliefs and those of others would change if we received unexpected information. The two kinds of reasoning interact, since one of the causal effects of a possible action open to me might be to give unexpected information to another rational agent.[2]

It is obvious that a possible course of events may be causally impossible even if it is epistemically open, as when you have already committed yourself, but I have not yet learned of your decision. It also may happen that a course of events is causally open even when it is epistemically closed in the sense that someone believes, with probability one, that it will not happen. But can it be true of a causally open course of events that someone not only believes, but also *knows* that it will not occur? This is less clear; it depends on how we understand the concept of knowledge. It does not seem incoherent to suppose that you know that I am rational, even though irrational choices are still causally possible for me. In fact, the concept of rationality seems applicable to actions only when there are options open to an agent. If we are to make sense of assumptions of knowledge and common knowledge of rationality, we need

[1]Ernest Adams (1970) first pointed to the contrast illustrated by this pair of conditionals. The particular example is Jonathan Bennett's.

[2]The relation between causal and evidential reasoning is the central concern in the development of causal decision theory. See Gibbard and Harper (1981), Skyrms (1982) and Lewis (1980).

to allow for the possibility that an agent may know what he or another agent is going to do, even when it remains true that the agent could have done otherwise.

To clarify the causal and epistemic concepts that interact in strategic reasoning, it is useful to break them down into their component parts. If, for example, there is a problem about exactly what it means to assume that there is common knowledge of rationality, it ought to be analyzed into problems about exactly what rationality is, or about what knowledge is, or about how common knowledge is defined in terms of knowledge. The framework I will use to represent these concepts is one that is designed to help reveal the compositional structure of such complex concepts: it is a formal semantic or model theoretic framework – specifically, the Kripkean 'possible worlds' framework for theorizing about modal, causal and epistemic concepts. I will start by sketching a simple conception of a model, in the model theorist's sense, of a strategic form game. Second, I will add to the simple conception of a model the resources to account for one kind of counterfactual reasoning, reasoning about belief revision. In these models we can represent concepts of rationality, belief and common belief, and so can define the complex concept of common belief in rationality, and some related complex concepts, in terms of their component parts. The next step is to consider the concept of knowledge, and the relation between knowledge and belief. I will look at some different assumptions about knowledge, and at the consequences of these different assumptions for the concepts of common knowledge and common knowledge of rationality. Then to illustrate the way some of the notions I discuss might be applied to clarify some counterfactual reasoning about games, I will discuss some familiar problems about backward induction arguments, using the model theory to sharpen the assumptions of those arguments, and to state and prove some theorems about the consequences of assumptions about common belief and knowledge.

Model Theory for Games

Before sketching the conception of a model of a game that I will be using, I will set out some assumptions that motivate it, assumptions that i think will be shared by most though not all, game theorists. First, I assume that a game is a *partial* description of a set or sequence of interdependent Bayesian decision problems. The description is partial in that while it specifies all the relevant utilities motivating the agents, it does not give their degrees of belief. Instead, qualitative constraints are put on what the agents are assumed to believe about the actions of other agents; but these constraints will not normally be enough to determine what the agents believe about each other, or to determine what solutions are prescribed to the decision problems. Second, I assume that all of the decision problems in the game are problems of individual decision making. There is no special concept of rationality for decision making in a situation where the outcomes depend on the actions of more than one agent. The acts of other agents are, like chance events, natural disasters and acts of God, just facts about an uncertain world that agents have beliefs and degrees of

belief about. The utilities of other agents are relevant to an agent only as information that, together with beliefs about the rationality of those agents, helps to predict their actions. Third, I assume that in cases where degrees of belief are undetermined, or only partially determined, by the description of a decision problem, then no action is prescribed by the theory unless there is an action that would be rational for every system of degrees of belief compatible with what is specified. There are no special rules of rationality telling one what to do in the absence of degrees of belief, except this: decide what you believe, and then maximize expected utility.

A model for a game is intended to represent a completion of the partial specification of the set or sequence of Bayesian decision problems that is given by the definition of the game, as well as a representation of a particular play of the game. The class of all models for a game will include all ways of filling in the relevant details that are compatible with the conditions imposed by the definition of the game. Although a model is intended to represent one particular playing of the game, a single model will contain many possible worlds, since we need a representation, not only of what actually happens in the situation being modeled, but also what might or would happen in alternative situations that are compatible with the capacities and beliefs of one or another of the agents. Along with a set of possible worlds, models will contain various relations and measures on the set that are intended to determine all the facts about the possible worlds that may be relevant to the actions of any of the agents playing the game in a particular concrete context.

The models considered in this paper are models for finite games in normal or strategic form. I assume, as usual, that the game Γ itself consists of a structure $\langle N, \langle C_i, u_i\rangle_{i\in N}\rangle$, where N is a finite set of players, C_i is a finite set of alternative strategies for player i, and u_i is player i′s utility function taking a strategy profile (a specification of a strategy for each player) into a utility value for the outcome that would result from that sequence of strategies. A model for a game will consist of a set of possible worlds (a state space), one of which is designated as the actual world of the model. In each possible world in the model, each player has certain beliefs and partial beliefs, and each player makes a certain strategy choice. The possible worlds themselves are simple, primitive elements of the model; the information about them – what the players believe and do in each possible world – is represented by several functions and relations given by a specification of the particular model. Specifically, a model for a game Γ will consist of a structure $\langle W, a, \langle S_i, R_i, P_i\rangle_{i\in N}\rangle$, where W is a nonempty set (the possible worlds), a is a member of W (the actual world), each S_i is a function taking possible worlds into strategy choices for player i, each R_i is a binary relation on W, and each P_i is an additive measure function on subsets of W.

The R relations represent the qualitative structure of the players' beliefs in the different possible worlds in the following way: the set of possible worlds that are compatible with what player i believes in world w is the set $\{x{:}wR_ix\}$. It is assumed

that the R relations are *serial, transitive,* and *euclidean*.[3] The first assumption is simply the requirement that in any possible world there must be at least one possible world compatible with what any player believes in that world. The other two constraints encode the assumption that players know their own minds: they are necessary and sufficient to ensure that players have introspective access to their beliefs: if they believe something, they believe that they believe it, and if they do not, they believe that they do not.

The S functions encode the facts about what the players do – what strategies they choose – in each possible world. It is assumed that if xR_iy, then $S_i(x) = S_i(y)$. Intuitively, this requirement is the assumption that players know, at the moment of choice, what they are doing – what choice they are making. Like the constraints on the structure of the R relations, this constraint is motivated by the assumption that players have introspective access to their own states of mind.

The measure function P_i, encodes the information about the player's *partial* beliefs in each possible world in the following way: player i′s belief function in possible world w is the relativization of P_i to the set $\{x{:}wR_ix\}$. That is, for any proposition ϕ, $P_{i,W}(\phi) = P_i(\phi \cap \{x{:}wR_ix\})/P_i(\{x{:}wR_ix\})$. The assumptions we are making about R_i and P_i will ensure that $P_i(\{x{:}wR_ix\})$ is nonzero for all w, so that this probability will always be defined.

The use of a single measure function for each player, defined on the whole space of possible worlds, to encode the information required to define the player's degrees of belief is just a technical convenience – an economical way to specify the many different belief functions that represent that player's beliefs in different possible worlds. No additional assumptions about the players' beliefs are implicit in this form of representation, since our introspection assumptions already imply that any two different belief states for a single player are disjoint, and any set of probability measures on disjoint sets can be represented by a single measure on the union of all the sets. This single measure will contain some extraneous information that has no representational significance – different total measures will determine the same set of belief functions – but this artifact of the model is harmless.[4]

[3]That is, for all players i $(x)(\exists y)xR_iy$, $(x)(y)(z)((xR_iy \ \& \ yR_iz) \rightarrow xR_iz)$, and $(x)(y)(z)((xR_iy \ \& \ xR_iz) \rightarrow yRiz)$.

[4]It has been suggested that there is a substantive, and implausible, assumption built into the way that degrees of belief are modeled: namely, that any two worlds in which a player has the same *full* beliefs he also has the same *partial* beliefs. But this assumption is a tautological consequence of the introspection assumption, which implies that a player fully believes that he himself has the partial beliefs that he in fact has. It does follow from the introspection assumptions that player j cannot be uncertain about player i's partial beliefs while being certain about all of i's full beliefs. But that is just because the totality of i's full beliefs includes his beliefs about his own partial beliefs, and by the introspection assumption, i's beliefs about his own partial beliefs are complete and correct. Nothing, however, prevents there being a model in which there are different worlds in which player i has full beliefs about objective facts that are exactly the same, even though the degrees of belief about such facts are different. This situation will be modeled by disjoint but isomorphic sets of possible worlds. In such a case, another player j might be certain about player i's full beliefs about everything except i's own partial beliefs, while being uncertain about i's partial beliefs.

In order to avoid complications that are not relevant to the conceptual issues I am interested in, I will be assuming throughout this discussion that our models are finite, and that the measure functions all assign nonzero probability to every nonempty subset of possible worlds.

We need to impose one additional constraint on our models, a constraint that is motivated by our concern with counterfactual reasoning. A specification of a game puts constraints on the causal consequences of the actions that may be chosen in the playing of the game, and we want these constraints to be represented in the models. Specifically, in a strategic form game, the assumption is that the strategies are chosen independently, which means that the choices made by one player cannot influence the beliefs or the actions of the other players. One could express the assumption by saying that certain counterfactual statements must be true in the possible worlds in the model: if a player had chosen a different strategy from the one he in fact chose, the other players would still have chosen the same strategies, and would have had the same beliefs, that they in fact had. The constraint we need to add is a closure condition on the set of possible worlds – a requirement that there be enough possible worlds of the right kind to represent these counterfactual possibilities.

For any world w and strategy s for player i, there is a world f(w,s) meeting the following four conditions:

1. for all $j \neq i$, if wR_jx, then $f(w,s)R_jx$.
2. if wR_ix, then $f(w,s)R_if(x,s)$.
3. $S_i(f(w,s)) = s$
4. $P_i(f(w,s)) = P_i(w)$.

Intuitively, f(w,s) represents the counterfactual possible world that, in w, is the world that would have been realized if player i, believing exactly what he believes in w about the other players, had chosen strategy s.

Any of the (finite) models constructed for the arguments given in this paper can be extended to (finite) models satisfying this closure condition. One simply adds, for each $w \in W$ and each strategy profile c, a world corresponding to the pair (w,c), and extending the R's, P's, and S's in a way that conforms to the four conditions.[5]

Because of our concern to represent counterfactual reasoning, it is essential that we allow for the possibility that players have false beliefs in some possible worlds, which means that a world in which they have certain beliefs need not itself be compatible with those beliefs. Because the epistemic structures we have defined allow for false belief, they are more general than the partition structures that will be more familiar to game theorists. An equivalence relation meets the three conditions we have imposed on our R relations, but in addition must be reflexive. To impose this

[5]More precisely, for any given model $M = \langle W,a,\langle S_i,R_i,P_i\rangle_{i\in N}\rangle$, not necessarily meeting the closure condition, define a new model M' as follows: $W' = W \times C$;$a' = \langle a,S(a)\rangle$; for all $w \in W$ and $c \in C$, $S'(\langle w,c\rangle) = c$; for all $x,y \in W$ and $c,d \in C$, $\langle x,c\rangle R'_i \langle y,d\rangle$ if the following three conditions are met: (i) xR_iy, (ii) $c_i = d_i$, and (iii) for all $j \neq i$, $S_j(y) = d_j$; $P'_i(\langle x,c\rangle) = P_i(x)$. This model will be finite if the original one was, and will satisfy the closure condition.

additional condition would be to assume that all players necessarily have only true beliefs. But even if an agent in fact has only true beliefs, counterfactual reasoning requires an agent to consider possible situations in which some beliefs are false. First, we want to consider belief contravening, or epistemic counterfactuals: how players would revise their beliefs were they to learn they were mistaken. Second, we want to consider deliberation which involves causal counterfactuals: a player considers what the consequences would be of his doing something he is not in fact going to do. In both cases, a player must consider possible situations in which either she or another player has a false belief.

Even though the R relations are not, in general, equivalence relations, there is a relation definable in terms of R that does determine a partition structure: say that two worlds x and y are *subjectively indistinguishable* for player i ($x \approx_i y$) if player i′s belief state in x is the same as it is in y. That is, $x \approx_i y$ if and only if $\{z{:}xR_iz\} = \{z{:}yR_iz\}$. Each equivalence class determined by a subjective indistinguishability relation will be divided into two parts: the worlds compatible with what the player believes, and the worlds that are not. In the regular partition models, all worlds are compatible with what the player believes in the world, and the two relations, R_i and $\approx_i$, will coincide.

To represent counterfactual reasoning, we must also allow for possible worlds in which players act irrationally. Even if I am resolved to act rationally, I may consider in deliberation what the consequences would be of acting in ways that are not. And even if I am certain that you will act rationally, I may consider how I would revise my beliefs if I learned that I was wrong about this. Even models satisfying some strong condition, such as common belief or knowledge that everyone is rational, will still be models that contain counterfactual possible worlds in which players have false beliefs, and worlds in which they fail to maximize expected utility.

The aim of this model theory is generality: to make, in the definition of a model, as few substantive assumptions as possible about the epistemic states and behavior of players of a game in order that substantive assumptions can be made explicit as conditions that distinguish some models from others. But of course the definition inevitably includes a range of idealizing and simplifying assumptions, made for a variety of reasons. Let me just mention a few of the assumptions that have been built into the conception of a model, and the reasons for doing so.

First, while we allow for irrational action and false belief, we do assume (as is usual) that players all have coherent beliefs that can be represented by a probability function on some nonempty space of possibilities. So in effect, we make the outrageously unrealistic assumption that players are logically omniscient. This assumption is made only because it is still unclear, either conceptually or technically, how to understand or represent the epistemic situations of agents that are not ideal in this sense. This is a serious problem, but not one I will try to address here.

Second, as I have said, it is assumed that players have introspective access to their beliefs. This assumption could be relaxed by imposing weaker conditions on the R relations, although doing so would raise both technical and conceptual problems. It is not clear how one acts on one's beliefs if one does not have introspective access to them. Some may object to the introspective assumption on the ground that a person

may have unconscious or inarticulate beliefs, but the assumption is not incompatible with this: if beliefs can be unconscious, so can beliefs about beliefs. It is not assumed that one knows how to say what one believes.

Third, some have questioned the assumption that players know what they do. This assumption might be relaxed with little effect; what is its motivation? The idea is simply that in a static model for a strategic form game, we are modeling the situation at the moment of choice, and it seems reasonable to assume that at that moment, the agent knows what choice is being made.

Fourth, it is assumed that players know the structure of the game – the options available and the utility values of outcomes for all of the players. This assumption is just a simplifying assumption made to avoid trying to do too much at once. It could easily be relaxed with minimal effect on the structure of the models, and without raising conceptual problems. That is, one could consider models in which different games were being played in different possible worlds, and in which players might be uncertain or mistaken about what the game was.

Finally, as noted we assume that models are finite. This is again just a simplifying assumption. Relaxing it would require some small modifications and add some mathematical complications, but would not change the basic story.

In any possible worlds model, one can identify *propositions* with subsets of the set of possible worlds, with what economists and statisticians call 'events'. The idea is to identify the content of what someone may think or say with, its truth conditions – that is, with the set of possible worlds that would realize the conditions that make what is said or thought true. For any proposition ϕ and player i, we can define the proposition that i fully believes that ϕ as the set $\{x \in W: \{y \in W: xR_i y\} \subseteq \phi\}$, and the proposition that i believes that ϕ to at least degree r as the set $\{x \in W: P_{i,x}(\phi) \geq r\}$. So we have the resources to interpret unlimited iterations of belief in any proposition, and the infinitely iterated concept of *common belief* (all players believe that ϕ, and all believe that all believe that ϕ, and all believe that all believe that all believe that ϕ, and ... etc.) can be defined as the intersection of all the propositions in this infinite conjunction. Equivalently, we can represent common belief in terms of the *transitive closure* R*, of the set all the R relations. For any proposition ϕ, it is, in possible world x, common belief among the players that ϕ if and only if ϕ is true in all possible worlds compatible with common belief, which is to say if and only if $\{y: xR^*y\} \subseteq \phi$.

If rationality is identified with maximizing expected utility, then we can define, in any model, the propositions that some particular player is rational, that all players are rational, that all players believe that all players are rational, and of course that it is common belief among the players that all players are rational. Here is a sequence of definitions, leading to a specification of the proposition that there is common belief that all players are rational[6]: first, the *expected utility* of an action (a strategy choice) s for a player i in a world x is defined in the familiar way:

[6]In these and other definitions, a variable for a strategy or profile, enclosed in brackets, denotes the proposition that the strategy or profile is realized. So, for example, if $e \in C_{-i}$ (if e is a strategy profile for players other than player i) then $[e] = \{x \in W: S_j(x) = e_j \text{ for all } j \neq i\}$.

$$\mathrm{eu}_{\mathrm{i,x}}(\mathrm{s}) = \sum_{\mathrm{e}\in \mathrm{C}_{-\mathrm{i}}} P_{i,x}([e]) \times u_i((s,\ e))$$

Second, we define the set of strategies that maximize expected utility for player i in world x:

$$\mathrm{r}_{\mathrm{i,x}} = \left\{\mathrm{s} \in \mathrm{C}_{\mathrm{i}}\ :\ \mathrm{eu}_{\mathrm{i,x}}(\mathrm{s}) \geq \mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}'\right) \text{ for all } \mathrm{s}' \in\ \mathrm{C}_{\mathrm{i}}\right\}$$

Third, the proposition *that player i is rational* is the set of possible worlds in which the strategy chosen maximizes expected utility in that world:

$$\mathrm{A}_{\mathrm{i}} = \{\mathrm{x} \in \mathrm{W}\ :\ \mathrm{S}_{\mathrm{i}}(\mathrm{x}) \in \mathrm{r}_{\mathrm{i,x}}\}$$

Fourth, the proposition *everyone is rational* is the intersection of the A_{i}'s:

$$\mathrm{A} = \cap_{\mathrm{i}\in \mathrm{N}} \mathrm{A}_{\mathrm{i}}$$

Fifth, the proposition *there is common belief that everyone is rational* is defined as follows:

$$\mathrm{Z} = \{\mathrm{x} \in \mathrm{W}\ :\ \{\mathrm{y} \in \mathrm{W} : \mathrm{xR}^*\mathrm{y}\} \subseteq \mathrm{A}\}.$$

Any specification that determines a proposition relative to a model can also be used to pick out a class of models – all the models in which the proposition is true in that model's actual world. So for any given game, we can pick out the class of models of that game that satisfy some intuitive condition, for example, the class of models in which the proposition Z, that there is common belief in rationality, is true (in the actual world of the model). A class of models defined this way in turn determines a set of strategy profiles for the game: a profile is a member of the set if and only if it is realized in the actual world of one of the models in the class of models. This fact gives us a way that is both precise and intuitively motivated of defining a solution concept for games, or of giving a proof of adequacy for a solution concept already defined. The solution concept that has the most transparent semantic motivation of this kind is rationalizability: we can define rationalizability semantically as the set of strategies of a game that are realized in (the actual world of) some model in which there is common belief in rationality.[7] Or, we can give

[7]This model theoretic definition of rationalizability coincides with the standard concept defined by Bernheim (1984) and Pearce (1984) only in two person games. In the general case, it coincides with the weaker concept, correlated rationalizability. Model theoretic conditions appropriate for the stronger definition would require that players' beliefs about each other satisfy a constraint that (in games with more than two players) goes beyond coherence: specifically, it is required that no player can believe that any information about another player's strategy choices would be evidentially relevant to the choices of a different player. I think this constraint could be motivated, in general, only if one confused causal with evidential reasoning. The structure of the game ensures

a direct nonsemantic definition of the set of strategies – the set of strategies that survive the iterated elimination of strictly dominated strategies – and then prove that this set is *characterized* by the class of models in which there is common belief in rationality: a set of strategies is characterized by a class of models if the set includes exactly the strategies that are realized in some model in the class.[8]

Belief Revision

There are many ways to modify and extend this simple conception of a model of a game. I will consider here just one embellishment, one that is relevant to our concern with counterfactual reasoning. This is the addition of some structure to model the players' policies for revising their beliefs in response to new information. We assume, as is usual, that rational players are disposed to revise their beliefs by conditionalization, but there is nothing in the models we have defined to say how players would revise their beliefs if they learned something that had a prior probability of 0 – something incompatible with the initial state of belief. A belief revision policy is a way of determining the sets of possible worlds that define the posterior belief states that would be induced by such information. The problem is not to generate such belief revision policies out of the models we already have – that is impossible. Rather, it is to say what new structure needs to be added to the model in order to represent belief revision policies, and what formal constraints the policies must obey.

Since we are modeling strategic form games, our models are static, and so there is no representation of any actual change in what is believed. But even in a static situation, one might ask how an agent's beliefs are disposed to change were he to learn that he was mistaken about something he believed with probability one, and the answer to this question may be relevant to his decisions. These dispositions to change beliefs, in contrast to the potential changes that would display the dispositions, are a part of the agent's prior subjective state – the only state represented in the worlds of our models.

I said at the start that one aim in constructing this model theory was to clarify, in isolation, the separate concepts that interact with each other in strategic contexts, and that are the component parts of the complex concepts used to describe those contexts. In keeping with this motivation, I will first look at a pure and simple abstract version of belief revision theory, for a single agent in a single possible

that players' strategy choices are made independently: if player one had chosen differently, it could not have influenced the choice of player two. But this assumption of causal independence has no consequences about the evidential relevance of information about player one's choice for the beliefs that a third party might rationally have about player two. (Brian Skyrms (1992, pp. 147–8) makes this point.)

[8]This characterization theorem is proved in Stalnaker (1994).

world, ignoring degrees of belief, and assuming nothing about the subject matter of the beliefs. After getting clear about the basic structure, I will say how to incorporate it into our models, with many agents, many possible worlds, and probability measures on both the prior and posterior belief states. The simple theory that I will sketch is a standard one that has been formulated in a number of essentially equivalent ways by different theorists.[9] Sometimes the theory is formulated syntactically, with prior and posterior belief states represented by sets of sentences of some formal language, but I will focus on a purely model theoretic formulation of the theory in which the agent's belief revision policy is represented by a set of possible worlds – the prior belief state – and a function taking each piece of potential new information into the conditional belief state that corresponds to the state that would be induced by receiving that information. Let B be the set representing the prior state, and let B′ be the set of all the possible worlds that are compatible with any new information that the agent could possibly receive. Then if ϕ is any proposition which is a subset of B′, B(ϕ) will be the set that represents the posterior belief state induced by information ϕ.

There are just four constraints that the standard belief revision theory imposes on this belief revision function:

1. For any ϕ, $B(\phi) \subseteq \phi$
2. If ϕ is nonempty, then B(ϕ) is nonempty
3. If $B \cap \phi$ is nonempty, then $B(\phi) = B \cap \phi$
4. If $B(\phi) \cap \psi$ is nonempty, then $B(\phi \& \psi) = B(\phi) \cap \psi$

The first condition is simply the requirement that the new information received is believed in the conditional state. The second is the requirement that consistent information results in a consistent conditional state. The third condition requires that belief change be conservative in the sense that one should not give up any beliefs unless the new information forces one to give something up: if ϕ is compatible with the prior beliefs, the conditional belief state will simply add ϕ to the prior beliefs. The fourth condition is a generalization of the conservative condition. Its effect is to require that if two pieces of information are received in succession, the second being compatible with the posterior state induced by the first, then the resulting change should be the same as if both pieces of information were received together.

Any belief revision function meeting these four conditions can be represented by an ordering of all the possible worlds, and any ordering of a set of possible worlds will determine a function meeting the four conditions. Let Q be any binary transitive and connected relation on a set B′. Then we can define B as the set of highest ranking members of B′, and for any subset ϕ of B′, we can define B(ϕ) as the set of highest ranking members of ϕ:

[9]The earliest formulation, so far as I know, of what has come to be called the AGM belief revision theory was given by William Harper (1975). For a general survey of the belief revision theory, see Gärdenfors (1988). Other important papers include Alchourón and Makinson (1982), Alchourón et al. (1985), Grove (1988), Makinson (1985) and Spohn (1987).

$$B(\phi) = \{x \in \phi \ : \ yQx \text{ for all } y \in \phi\}$$

It is easy to show that this function will satisfy the four conditions. On the other hand, given any revision function meeting the four conditions, we can define a binary relation Q in terms of it as follows:

$$xQy \text{ if } y \in B(\{x, y\}).$$

It is easy to show, using the four conditions, that Q, defined this way, is transitive and connected, and that $B(\phi) = \{x \in \phi$: yQx for all $y \in \phi\}$. So the specification of such a Q relation is just an alternative formulation of the same revision theory.

Now to incorporate this belief revision theory into our models, we need to give each player such a belief revision policy in each possible world. This will be accomplished if we add to the model a binary relation Q for each player. We need just one such relation for each player, if we take our assumption that players know their own states of mind to apply to belief revision policies as well as to beliefs themselves. Since the belief revision policy is a feature of the agent's subjective state, it is reasonable to assume that in all possible worlds that are subjectively indistinguishable for a player, he has the same belief revision policies.

Subjective indistinguishability (which we defined as follows: $x \approx_i y$ if and only if $\{z{:}xR_iz\} = \{z{:}yR_iz\}$) is an equivalence relation that partitions the space of all possible worlds for each player, and the player's belief revision function will be the same for each world in the equivalence class. (The equivalence class plays the role of B' in the simple belief revision structure.) What we need to add to the game model is a relation Q_i for each player that orders all the worlds within each equivalence class with respect to epistemic plausibility, with worlds compatible with what the player believes in the worlds in that class having maximum plausibility. So Q_i must meet the following three conditions:

(q1) $x \approx_i y$, if and only if xQ_iy or yQ_ix.
(q2) Q_i is transitive.
(q3) xR_iy if and only if wQ_iy for all w such that $w \approx_i x$.

For any proposition ϕ, we can define the conditional belief state for player i in world x, $B_{i,x}(\phi)$ (the posterior belief state that would be induced by learning ϕ),[10] in terms of Q_i as follows:

$$B_{i,x}(\phi) = \{w \in \phi : \text{for all } y \in \phi \cap \{z : z \approx_i x\}, yQ_iw\}.$$

[10]There is this difference between the conditional belief state $B_{i,x}(\phi)$ and the posterior belief state that would actually result if the agent were in fact to learn that ϕ: if he were to learn that ϕ, he would believe that he *then* believed that ϕ, whereas in our static models, there is no representation of what the agent comes to believe in the different possible worlds at some later time. But the potential posterior belief states and the conditional belief states as defined do not differ with respect to any information represented in the model. In particular, the conditional and posterior belief states do not differ with respect to the agent's beliefs about his *prior* beliefs.

Once we have added to our models a relation Q_i for each player that meets these three conditions, the R relations become redundant, since they are definable in terms of Q.[11] For a more economical formulation of the theory, we drop the R_i′s when we add the Q_i′s, taking condition (q1) as above the new definition of subjective indistinguishability, and condition (q3) as the definition of R_i. Formulated this way, the models are now defined as follows:

A model is a structure $<W,a,<S_i,Q_i,P_i>_{i\in N}>$. W, a, P_i, and S_i are as before; Each Q_i is a binary reflexive transitive relation on W meeting in addition the following condition: any two worlds that are Q_i related (in either direction) to a third world are Q_i related (in at least one direction) to each other. One can then prove that each R_i, defined as above, is serial, transitive, and euclidean. So our new models incorporate and refine models of the simpler kind.

To summarize, the new structure we have added to our models expresses exactly the following two assumptions:

1. In each possible world each player has a belief revision policy that conforms to the conditions of the simple AGM belief revision theory sketched above, where (for player i and world x) the set B is $\{y:xR_iy\}$, and the set B′ is $\{y: y \approx_i x\}$
2. In each world, each player has a correct belief about what his own belief revision policy is.

Each player's belief revision structure determines a ranking of all possible worlds with respect to the player's degree of epistemic success or failure in that world. In some worlds, the player has only true beliefs; in others, he makes an error, but not as serious an error as he makes in still other possible worlds. Suppose I am fifth on a standby waiting list for a seat on a plane. I learn that there is only one unclaimed seat, and as a result I feel certain that I will not get on the plane. I believe that the person at the top of the list will certainly take the seat, and if she does not, then I am certain that the second in line will take it, and so on. Now suppose in fact that my beliefs are mistaken: the person at the top of the list turns the seat down, and the next person takes it. Then my initial beliefs were in error, but not as seriously as they would be if I were to get the seat. If number two gets the seat, then I was making a simple first degree error, while if I get the seat, I was making a fourth degree error.

It will be useful to define, recursively, a sequence of propositions that distinguish the possible worlds in which a player's beliefs are in error to different degrees:

[11] The work done by Q is to rank the worlds incompatible with prior beliefs; it does not distinguish between worlds compatible with prior beliefs – they are ranked together at the top of the ordering determined by Q. So Q encodes the information about what the prior beliefs are – that is why R becomes redundant. A model with both Q and R relations would specify the prior belief sets in two ways. Condition (q3) is the requirement that the two specifications yield the same results.

Here is a simple abstract example, just to illustrate the structure: suppose there are just three possible worlds, x y and z, that are subjectively indistinguishable in those worlds to player i. Suppose {x} is the set of worlds compatible with i′s beliefs in x, y, and z, which is to say that the R relation is the following set: {<x,x>,<y,x>,<z,x>}. Suppose further that y has priority over z, which is to say if i were to learn the proposition {y,z}, his posterior or conditional belief state would be {y}. In other words, the Q relation is the following set: {<x,x>,<y,x>,<z,x>,<y,y>,<z,y>,<z,z>}.

E^1_i is the proposition that player i has at least some false belief – makes at least a simple first degree error.

$E^1_i = \{x \in W$: for some y such that $y \approx_i x$, not $yQ_ix\}(= \{x \in W$: not $xR_ix\})$

E^{k+1}_i is the proposition that player i makes at least a $k+1$ degree error:

$$E^{k+1}_i = \left\{x \in E^k_i \ : \text{ for some } y \in E^k_i \text{ such that } y \approx_i x, \text{ not } yQ_ix\right\}.$$

The belief revision structure provides for epistemic distinctions between propositions that are all believed with probability one. Even though each of two propositions has maximum degree of belief, one may be believed *more robustly* than the other in the sense that the agent is more disposed to continue believing it in response to new information. Suppose, to take a fanciful example, there are three presidential candidates, George, a Republican from Texas, Bill, a Democrat from Arkansas, and Ross, an independent from Texas. Suppose an agent believes, with probability one, that George will win. She also believes, with probability one, that a Texan will win and that a major party candidate will win, since these follow, given her other beliefs, from the proposition that George will win. But one of these two weaker beliefs may be more robust than the other. Suppose the agent is disposed, on learning that George lost, to conclude that Bill must then be the winner. In this case, the belief that a major party candidate will win is more robust than the belief that a Texan will win.

The belief revision structure is purely qualitative, but the measure functions that were already a part of the models provide a measure of the partial beliefs for conditional as well as for prior belief states. The Q relations, like the R relations, deliver the sets of possible worlds relative to which degrees of belief are defined. The partial beliefs for conditional belief state, like those for the prior states, are given by relativizing the measure function to the relevant set of possible worlds. Just as player i's partial beliefs in possible world x are given by relativizing the measure to the set $B_{i,x} = \{y: xR_iy\}$, so the partial beliefs in the conditional belief state for player i, world x and condition ϕ is given by relativizing the measure to the set $B_{i,x}(\phi) = \{y \in \phi$: for all $z \in \phi$ such that $z \approx_i x$, $zQ_iy\}$.

So with the help of the belief revision function we can define *conditional* probability functions for each player in each world:

$$P_{i,x}(\phi/\psi) = P_i(\phi \cap B_{i,x}(\psi))/P_i(B_{i,x}(\psi)$$

In the case where the condition ψ is compatible with i′s prior beliefs – where $P_{i,x}(\psi) > 0$ – this will coincide with conditional probability as ordinarily defined. (This is ensured by the conservative condition on the belief revision function.) But this definition extends the conditional probability functions for player x in world i to any condition compatible with the set of worlds that are subjectively indistinguishable for x in i.[12]

[12]These extended probability functions are equivalent to *lexicographic probability systems.* See Blume et al. (1991a, b) for an axiomatic treatment of lexicographic probability in the context of decision theory and game theory. These papers discuss a concept equivalent to the one defined below that I am calling perfect rationality.

The belief revision theory, and the extended probability functions give us the resources to introduce a refinement of the concept of rationality. Say that an action is *perfectly rational* if it not only maximizes expected utility, but also satisfies a tie-breaking procedure that requires that certain *conditional* expected utilities be maximized as well. The idea is that in cases where two or more actions maximize expected utility, the agent should consider, in choosing between them, how he should act if he learned he was in error about something. And if two actions are still tied, the tie-breaking procedure is iterated – the agent considers how he should act if he learned that he were making an error of a higher degree. Here is a sequence of definitions leading to a definition of perfect rationality.

Given the extended conditional probability functions, the definition of conditional expected utility is straightforward:

$$\mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}/\phi\right) = \sum_{e \in \mathrm{C}_{-\mathrm{i}}} P_{i,x}\left([e]/\phi\right) \times u_i\left((s,e)\right)$$

Second, we define, recursively, a sequence of sets of strategies that maximize expected utility, and also satisfy the succession of tie-breaking rules:

$$\mathrm{r}^{0}_{\mathrm{i,x}} = \mathrm{r}_{\mathrm{i,x}}\left(\text{that is, } \left\{\mathrm{s} \in \mathrm{C}_{\mathrm{i}} : \mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}\right) \geq \mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}'\right) \text{ for all } \mathrm{s}' \in \mathrm{C}_{\mathrm{i}}\right\}\right)$$

$$\mathrm{r}^{\mathrm{k}+1}_{\mathrm{i,x}} = \left\{\mathrm{s} \in \mathrm{r}^{\mathrm{k}}_{\mathrm{i,x}} : \mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}/\mathrm{E}^{\mathrm{k}+1}\right) \geq \mathrm{eu}_{\mathrm{i,x}}\left(\mathrm{s}'/\mathrm{E}^{\mathrm{k}+1}\right) \text{ for all } \mathrm{s}' \in \mathrm{r}^{\mathrm{k}}_{\mathrm{i,x}}\right\}.$$

$$\mathrm{r}^{+}_{\mathrm{i,x}} = \cap \mathrm{r}^{\mathrm{k}}_{\mathrm{i,x}} \text{ for all k such that } \mathrm{E}^{\mathrm{k}} \cap \{\mathrm{y} : \mathrm{x} \approx_{\mathrm{i}} \mathrm{y}\} \text{ is nonempty.}$$

The set $\mathrm{r}^{+}_{\mathrm{i,x}}$ is the set of strategies that are perfectly rational for player i in world x. So the proposition that player i is perfectly rational is defined as follows:

$$\mathrm{A}^{+}_{\mathrm{i}} = \left\{\mathrm{x} \in \mathrm{W} : \mathrm{S}_{\mathrm{i}}\left(\mathrm{x}\right) \in \mathrm{r}^{+}_{\mathrm{i,x}}\right\}.$$

I want to emphasize that this refinement is defined wholly within individual decision theory. The belief revision theory that we have imported into our models is a general, abstract structure, as appropriate for a single agent facing a decision problem to which the actions of other agents are irrelevant as it is for a situation in which there are multiple agents. It is sometimes said that while states with probability 0 are

I don't want to suggest that this is the only way of combining the AGM belief revision structure with probabilities. For a very different kind of theory, see Mongin (1994). In this construction, probabilities are nonadditive, and are used to represent the belief revision structure, rather than to supplement it as in the models I have defined. I don't think the central result in Mongin (1994) (that the same belief revision structure that I am using is in a sense equivalent to a nonadditive, and so non-Bayesian, probability conception of prior belief) conflicts with, or presents a problem for, the way I have defined extended probability functions: the probability numbers just mean different things in the two constructions.

relevant in game theory, they are irrelevant to individual decision making,[13] but I see no reason to make this distinction. There is as much or as little reason to take account, in one's deliberation, of the possibility that nature may surprise one as there is to take account of the possibility that one may be fooled by one's fellow creatures.

Perfect rationality is a concept of individual decision theory, but in the game model context this concept may be used to give a model theoretic definition of a refinement of rationalizability. Say that a strategy of a game Γ is *perfectly rationalizable* if and only if the strategy is played in some model of Γ in which the players have common belief that they all are perfectly rational. As with ordinary correlated rationalizability, one can use a simple algorithm to pick out the relevant class of strategies, and prove a characterization theorem that states that the model theoretic and algorithmic definitions determine the same class of strategies. Here is the theorem:

Strategies that survive the elimination of all weakly dominated strategies followed by the iterated elimination of strictly dominated strategies are all and only those that are realized in a model in which players have common belief that all are perfectly rational.[14]

Before going on to discuss knowledge, let me give two examples of games to illustrate the concepts of perfect rationality and perfect rationalizability.

First, consider the following very simple game: Alice can take a dollar for herself alone, ending the game, or instead leave the decision up to Bert, who can either decide whether the two players get a dollar each, or whether neither gets anything. Figure 42.1 represents the strategic form of this game.

Both strategies for both players are rationalizable, but only Tt is perfectly rationalizable. If Alice is certain that Bert will play t, then either of her strategies would maximize expected utility. But only choice T will ensure that utility is maximized also on the condition that her belief about Bert's choice is mistaken. Similarly, Bert may be certain that Alice won't give him the chance to choose, but if he has to commit himself to a strategy in advance, then if he is perfectly rational, he will opt for the choice that would maximize expected utility if he did get a chance to choose.

[13]For example, Fudenberg and Tirole (1992) make the following remark about the relation between game theory and decision theory: 'Games and decisions differ in one key respect: probability-0 events are both exogenous and irrelevant in decision problems, whereas what *would* happen if a player played differently in a game is both important and endogenously determined'.

To the extent that this is true, it seems to me an accident of the way the contrasting theories are formulated, and to have no basis in any difference in the phenomena that the theories are about.

[14]The proof of this theorem, and others stated without proof in this paper, are available from the author. The argument is a variation of the proof of the characterization theorem for simple (correlated) rationalizability given in Stalnaker (1994). See Dekel and Fudenberg (1990) for justification of the same solution concept in terms of different conditions that involve perturbations of the payoffs.

I originally thought that the set of strategies picked out by this concept of perfect rationalizability coincided, in the case of two person games, with perfect rationalizability as defined by Bernheim (1984), but Pierpaolo Battigalli pointed out to me that Bernheim's concept is stronger.

Fig. 42.1

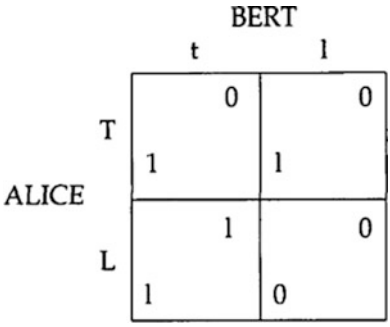
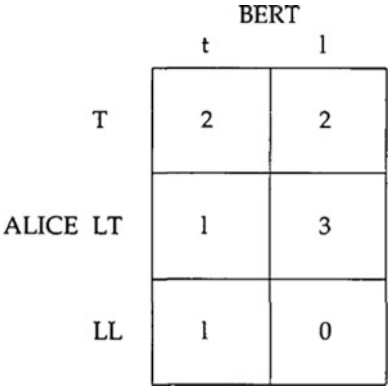


Fig. 42.2



Second, consider the following pure common interest game, where the only problem is one of coordination. It is also a perfect information game. One might think that coordination is no problem in a perfect information game, but this example shows that this is not necessarily true.

Alice can decide that each player gets two dollars, ending the game, or can leave the decision to Bert, who may decide that each player get one dollar, or may give the decision back to Alice. This time, Alice must decide whether each player gets three dollars, or neither gets anything. Figure 42.2 represents the strategic form of this game.

Now suppose Bert believes, with probability one, that Alice will choose T; what should he do? This depends on what he thinks Alice would do on the hypothesis that his belief about her is mistaken. Suppose that, if he were to be surprised by Alice choosing L on the first move, he would conclude that, contrary to what he previously believed, she is irrational, and is more likely to choose L on her second choice as well. Given these belief revision policies, only choice t is perfectly rational for him. But why should Alice choose T? Suppose she is sure that Bert will choose t, which as we have just seen, is the only perfectly rational choice for him to make if his beliefs about Alice are as we have described. Then Alice's only rational choice is T. So it might be that Alice and Bert both know each others' beliefs about each other, and are both perfectly rational, but they still fail to coordinate on the optimal

outcome for both. Of course nothing in the game requires that Bert and Alice should have these beliefs and belief revision policies, but the game is compatible with them, and with the assumption that both Bert and Alice are perfectly rational.

Now one might be inclined to question whether Bert really believes that Alice is fully rational, since he believes she would choose L on her second move, if she got a second move, and this choice, being strictly dominated, would be irrational. Perhaps if Bert believed that Alice was actually disposed to choose L on her second move, then he wouldn't believe she was fully rational, but it is not suggested that he believes this. Suppose we divide Alice's strategy T into two strategies, TT and TL, that differ only in Alice's counterfactual dispositions: the two strategies are 'T, and I *would* choose T again on the second move if I were faced with that choice', and 'T, but I *would* choose L on the second move if I were faced with that choice'. One might argue that only TT, of these two, could be fully rational, but we may suppose that Bert believes, with probability one, that Alice will choose TT, and not TL. But were he to learn that he is wrong – that she did not choose TT (since she did not choose T on the first move) he would conclude that she instead chooses LL. To think there is something incoherent about this combination of beliefs and belief revision policy is to confuse epistemic with causal counterfactuals – it would be like thinking that because I believe that if Shakespeare hadn't written Hamlet, it would have never been written by anyone, I must therefore be disposed to conclude that Hamlet was never written, were I to learn that Shakespeare was in fact not its author.

Knowledge

As has often been noted, rationalizability is a very weak constraint on strategy choice, and perfect rationalizability is only slightly more restrictive. Would it make any difference if we assumed, not just common *belief* in rationality, or perfect rationality, but common *knowledge* as well? Whether it makes a difference, and what difference it makes, will depend on how knowledge is analyzed, and on what is assumed about the relation between knowledge and belief. I will consider a certain analysis of knowledge with roots in the philosophical literature about the definition of knowledge, an analysis that can be made precise with the resources of the belief revision structure that we have built into our models. But before getting to that analysis, I want to make some general remarks about the relation between knowledge and belief.

Whatever the details of one's analysis of knowledge and belief, it is clear that the central difference between the two concepts is that the first, unlike the second, can apply only when the agent is in fact correct in what he believes: the claim that i knows that ϕ, in contrast with the claim that i believes that ϕ, entails that ϕ, is true. Everyone knows that knowledge is different from belief – even from the extreme of belief, probability one – in this way, but sometimes it is suggested that this difference does not matter for the purposes of decision theory, since the rationality of a decision is independent of whether the beliefs on which it is based are in fact correct. It is

expected utility, not the value of the actual payoff that I receive in the end, that is relevant to the explanation and evaluation of my actions, and expected utility cannot be influenced by facts about the actual world that do not affect my beliefs. But as soon as we start looking at one person's beliefs and knowledge about another's beliefs and knowledge, the difference between the two notions begins to matter. The assumption that Alice believes (with probability one) that Bert believes (with probability one) that the cat ate the canary tells us nothing about what Alice believes about the cat and the canary themselves. But if we assume instead that Alice *knows* that Bert *knows* that the cat ate the canary, it follows, not only that the cat in fact ate the canary, but that Alice knows it, and therefore believes it as well.

Since knowledge and belief have different properties, a concept that conflates them will have properties that are appropriate for neither of the two concepts taken separately. Because belief is a subjective concept, it is reasonable to assume, as we have, that agents have introspective access to what they believe, and to what they do not believe. But if we switch from belief to knowledge, an external condition on the cognitive state is imposed, and because of this the assumption of introspective access is no longer tenable, even for logically omniscient perfect reasoners whose mental states are accessible to them. Suppose Alice believes, with complete conviction and with good reason that the cat ate the canary, but is, through no fault of her own, factually mistaken. She *believes,* let us suppose, that she knows that the cat ate the canary, but her belief that she knows it cannot be correct. Obviously, no amount of introspection into the state of her own mind will reveal to her the fact that she lacks this knowledge. If we conflate knowledge and belief, assuming in general that i knows that ϕ if and only if i′s degree of belief for ϕ is one, then we get a concept that combines the introspective properties appropriate only to the internal, subjective concept of belief with the success properties appropriate only to an external concept that makes claims about the objective world. The result is a concept of knowledge that rests on equivocation.

The result of this equivocation is a concept of knowledge with the familiar partition structure, the structure often assumed in discussions by economists and theoretical computer scientists about common knowledge, and this simple and elegant structure has led to many interesting results.[15] But the assumption that knowledge and common knowledge have this structure is the assumption that there can be no such thing as false belief, that while ignorance is possible, error is not. And since there is no false belief, there can be no disagreement, no surprises, and no coherent counterfactual reasoning.[16]

[15]Most notably, Robert Aumann's important and influential result on the impossibility of agreeing to disagree, and subsequent variations on it all depend on the partition structure, which requires the identification of knowledge with belief. See Aumann (1976) and Bacharach (1985). The initial result is striking, but perhaps slightly less striking when one recognizes that the assumption that there is no disagreement is implicitly a premise of the argument.

[16]If one were to add to the models we have defined the assumption that the R relation is reflexive, and so (given the other assumptions) is an equivalence relation, the result would be that the three relations, R_i, Q_i, and $\approx_i$, would all collapse into one. There would be no room for belief revision,

It is sometimes suggested that if one were to analyze knowledge simply as true belief, then the result would be a concept of knowledge with this partition structure, but this is not correct. The conjunctive concept, true belief, will *never* determine a partition structure unless it is assumed that it is necessary that *all* beliefs are true, in which case the conjunctive concept would be redundant. For suppose there might be a false belief – that it might be that some person i believed that ϕ, but was mistaken. Then it is false that i truly believes that ϕ, and so if true belief satisfied the conditions of the partition structure, it would follows that i truly believes that he does not truly believe that ϕ, from which (since he believes ϕ) he could infer that ϕ is false. The point is that to assume negative introspection for true belief is to assume that a believer can distinguish, introspecrively, her true beliefs from her false beliefs, which implies (at least if she is consistent) that she won't have any false beliefs.

While it can never be reasonable to equate knowledge and belief in general, we can specify the contingent conditions under which knowledge and belief will coincide. If we assume about a particular situation that *as a matter of fact,* a person has no false beliefs, then (and only then) can we conclude that in that situation, knowledge and belief coincide. To get this conclusion, we need to make no assumptions about knowledge beyond the minimal one that knowledge implies true belief. The assumption we need to make is that full belief is a state that is subjectively indistinguishable from knowledge: that fully believing that ϕ is the same as fully believing that one knows that ϕ.

If we make the idealizing assumption about a particular game situation being modeled that no one has any false beliefs, and that it is common belief that no one has any false beliefs, then we can have the benefits of the identification of knowledge and belief without the pernicious consequences that come from equivocating between them. What we cannot and need not assume is that it is a *necessary* truth – true in all possible worlds in the model – that no one has any false beliefs. Even if players *actually* have only true beliefs, there will inevitably be *counterfactual* possible worlds in the model in which players have false beliefs. These counterfactual possible worlds must be there to represent the causal possibilities that define the structure of the game, and to represent the belief revision policies of the players. If we assumed that it was a necessary truth that there was no false belief, then it would be impossible for one player to believe that a second player was rational in any model for any game in which irrational options are available to the second player.

In terms of this idealizing assumption about knowledge and belief, we can define a refinement of rationalizability, which I have called *strong rationalizability*. Here

since it would be assumed that no one had a belief that could be revised. Intuitively, the assumption would be that it is a necessary truth that all players are Cartesian skeptics: they have no probability-one beliefs about anything except necessary truths and facts about their own states of mind. This assumption is not compatible with belief that another player is rational, unless it is assumed that it is a necessary truth that the player is rational.

is the model theoretic definition: for any game Γ a strategy profile is strongly rationalizable if and only if it is realized in a model in which there is no error, common belief that all players are rational, and common belief that there is no error. The set of strategy profiles characterized by this condition can also be given an algorithmic definition, using an iterated elimination procedure intermediate between the elimination of strictly dominated and of weakly dominated strategies.[17] We can also define a further refinement, *strong perfect rationalizability*: just substitute 'perfect rationality' for 'rationality' in the condition defining strong rationalizability. A minor variation of the algorithm will pick out the set of strategy profiles characterized by these conditions.

Knowledge and belief coincide on this demanding idealization, but suppose we want to consider the more general case in which a person may know some things about the world, even while being mistaken about others. How should knowledge be analyzed? The conception of knowledge that I will propose for consideration is a simple version of what has been called, in the philosophical literature about the analysis of knowledge, the *defeasibility analysis*. The intuitive idea behind this account is that 'if a person has knowledge, then that person's justification must be sufficiently strong that it is not capable of being *defeated* by evidence that he does not possess' (Pappas and Swain 1978). According to this idea, if evidence that is unavailable to you would give you reason to give up a belief that you have, then your belief rests in part on your ignorance of that evidence, and so even if that belief is true, it will not count as knowledge.

We can make this idea precise by exploiting the belief revision structure sketched above, and the notion of robustness that allowed us to make epistemic distinctions between propositions believed with probability one. The analysis is simple: i knows that ϕ if and only if i believes that ϕ (with probability one), *and that belief is robust with respect to the truth*. That is, i knows that ϕ in a possible world x if and only if ϕ receives probability one from i in x, and also receives probability one in every conditional belief state for which the condition is true in x. More precisely, the proposition that i knows that ϕ is the set $\{x \in W$: for all ψ such that $x \in \psi$, $B_{i,x}(\psi) \subseteq \phi\}$.

Let me illustrate the idea with the example discussed above of the presidential candidates. Recall that there are three candidates, George, Bill and Ross, and that the subject believes, with probability one, that George will win. As a result she also believes with probability one that a Texan will win, and that a major party candidate will win. But the belief that a major party candidate will win is more robust than the belief that a Texan will win, since our subject is disposed, should she learn that George did not win, to infer that the winner was Bill. Now suppose, to everyone's surprise, Ross wins. Then even though our subject's belief that a Texan would win turned out to be true, it does not seem reasonable to say that she *knew* that a Texan would win, since she was right only by luck. Had she known more (that George

[17]The algorithm, which eliminates iteratively profiles rather than strategies, is given in Stalnaker (1994), and it is also proved there that the set of strategies picked out by this algorithm is characterized by the class of models meeting the model theoretic condition.

would lose), then that information would have undercut her belief. On the other hand, if Bill turns out to be the winner, then it would not be unreasonable to say that she knew that a major party candidate would win, since in this case her belief did not depend on her belief that it was George rather than Bill that would win.

The defeasibility conception of knowledge can be given a much simpler definition in terms of the belief revision structure. It can be shown that the definition given above is equivalent to the following: the proposition *i knows that* ϕ is the set $\{x: \{y:xQ_iy\} \subseteq \phi\}$. This exactly parallels the definition of the proposition that i believes that ϕ: $\{x: \{y:xR_iy\} \subseteq \phi\}$. On the defeasibility analysis, the relations that define the belief revision structure are exactly the same as the relations of epistemic accessibility in the standard semantics for epistemic logic.[18] And common knowledge (the infinite conjunction, everyone knows that ϕ, everyone knows that everyone knows that ϕ, ...) exactly parallels common belief: the proposition *there is common knowledge that* ϕ is $\{x:\{y:xQ^*y\} \subseteq \phi\}$, where Q^* is the transitive closure of the Q_i relations.

The defeasibility analysis provides us with two new model theoretic conditions that can be used to define solution concepts: first, the condition that there is common knowledge of rationality; second, the condition that there is common knowledge of perfect rationality. The conditions are stronger (respectively) than the conditions we have used to characterize rationalizability and perfect rationalizability, but weaker than the conditions that characterize the concepts I have called strong rationalizability and strong perfect rationalizability. That is, the class of models in which there is common belief in (perfect) rationality properly includes the class in which there is common knowledge, in the defeasibility sense, of (perfect) rationality, which in turn properly includes the class in which there is no error, common belief that there is no error, and common belief in (perfect) rationality. So the defeasibility analysis gives us two distinctive model theoretic solution concepts, but surprisingly, the sets of strategy profiles characterized by these new model theoretic conditions are the same as those characterized, in one case, by the weaker condition, and in the other case by the stronger condition. That is, the following two claims are theorems:

1. *Any strategy realized in a model in which there is common belief in (simple) rationality is also realized in a model in which there is common knowledge (in the defeasibility sense) of rationality.*
2. *Any strategy profile realized in a model in which there is common knowledge of perfect rationality is also realized in a model meeting in addition the stronger condition that there is common belief that no one has a false belief.*[19]

[18]The modal logic for the knowledge operators in a language that was interpreted relative to this semantic structure would be S4.3. This is the logic characterized by the class of Kripke models in which the accessibility relation is transitive, reflexive, and weakly connected (if xQ_iy and xQ_iz, then either yQ_iz or zQ_iy). The logic of common knowledge would be S4.

[19]Each theorem claims that any strategy that is realized in a model of one kind is also realized in a model that meets more restrictive conditions. In each case the proof is given by showing how to modify a model meeting the weaker conditions so that it also meets the more restrictive conditions.

Backward Induction

To illustrate how some of this apparatus might be deployed to help clarify the role in strategic arguments of assumptions about knowledge, belief and counterfactual reasoning, I will conclude by looking at a puzzle about backward induction reasoning, focusing on one notorious example: the finite iterated prisoners' dilemma. The backward induction argument purports to show that if there is common belief, or perhaps common knowledge, that both players are rational, then both players will defect every time, from the beginning. Obviously rational players will defect on the last move, and since they know this on the next to last move, they will defect then as well, and so on back through the game. This kind of argument is widely thought to be paradoxical, but there is little agreement about what the paradox consists in. Some say that the argument is fallacious, others that it shows an incoherence in the assumption of common knowledge of rationality, and still others that it reveals a self-referential paradox akin to semantic paradoxes such as the liar. The model theoretic apparatus we have been discussing gives us the resources to make precise the theses that alternative versions of the argument purport to prove, and to assess the validity of the arguments. Some versions are clearly fallacious, but others, as I will show, are valid.

The intuitive backward induction argument applies directly to games in extensive form, whereas our game models are models of static strategic form games.[20] But any extensive form game has a unique strategic form, and proofs based on the idea of the intuitive backward induction argument can be used to establish claims about the strategic form of the game. A backward induction argument is best seen as an argument by mathematical induction about a class of games that is closed with respect to the subgame relation – in the case at hand, the class of iterated prisoners' dilemmas of length n for any natural number n.

The conclusions of the backward induction arguments are conditional theses: if certain conditions obtain, then players will choose strategies that result in defection every time. The conditions assumed will correspond to the constraints on models that we have used to characterize various solution concepts, so the theses in question will be claims that only strategy profiles that result in defection every time will satisfy the conditions defining some solution concept. If, for example, the conditions are that there is common belief in rationality, then the thesis would be that only strategies that result in defection every time are rationalizable. It is clear that a backward induction argument for this thesis must be fallacious since many

[20] Although in this paper we have considered only static games, it is a straightforward matter to enrich the models by adding a temporal dimension to the possible worlds, assuming that players have belief states and perform actions at different times, actually revising their beliefs in the course of the playing of the game in accordance with a belief revision policy of the kind we have supposed. Questions about the relationship between the normal and extensive forms of games, and about the relations between different extensive-form games with the same normal form can be made precise in the model theory, and answered.

cooperative strategies are rationalizable. Pettit and Sugden (1989) have given a nice diagnosis of the fallacy in this version of the argument. But what if we make the stronger assumption that there is common *knowledge* of rationality, or of perfect rationality? Suppose, first, that we make the idealizing assumption necessary for identifying knowledge with belief: that there is no error and common belief that there is no error, and common belief that both players are rational. Are all *strongly* rationalizable strategy pairs in the iterated prisoners' dilemma pairs that result in defection every time? In this case the answer is positive, and the theorem that states this conclusion is proved by a backward induction argument.

To prove this backward induction theorem, we must first prove a lemma that is a general claim about multi-stage games – a class of games that includes iterated games. First, some notation and terminology: let Γ be any game that can be represented as a multi-stage game with observed action (a game that can be divided into stages where at each stage all players move simultaneously, and all players know the result of all previous moves). Let $\Gamma^{\#}$ be any subgame – any game that begins at the start of some later stage of Γ. For any strategy profile c of Γ that determines a path through the subgame $\Gamma^{\#}$, let $c^{\#}$ be the profile for $\Gamma^{\#}$ that is determined by c, and let $C^{\#}$ be the set of all strategy profiles of Γ that determine a path through $\Gamma^{\#}$. By 'an SR model', I will mean a model in which there is (in the actual world of the model) no error, common belief that there is no error, and common belief that all players are rational. Now we can state the multi-stage game lemma:

If profile c is strongly rationalizable in Γ, and if c determines a path through $\Gamma^{\#}$, then $c^{\#}$ is strongly rationalizable in $\Gamma^{\#}$.

This is proved by constructing a model for $\Gamma^{\#}$ in terms of a model for Γ, and showing that if the original model is an SR model, so is the new one. Let M be any SR model for Γ in which c is played in the actual world of the model. Let $\Gamma^{\#}$ be any subgame that contains the path determined by c. We define a model $M^{\#}$ for $\Gamma^{\#}$ in terms of M as follows: $W^{\#} = \{x \in W{:}S(x) \in C^{\#}\}$. The $Q_i{}^{\#}$'s and $P_i{}^{\#}$'s are simply the restrictions of the Q_i's and P_i's to $W^{\#}$. The $S_i{}^{\#}$'s are defined so that for each $x \in W^{\#}$, $S^{\#}(x)$ is the profile for the game $\Gamma^{\#}$ that is determined by the profile S(x). (That is, if $S(x) = e$, then $S^{\#}(x) = e^{\#}$.)

To see that $M^{\#}$ is an SR model for $\Gamma^{\#}$, note first that if there is no error and common belief that there is no error in the original model, then this will also hold for the model of the subgame: if $\{x{:}aR^{*}x\} \subseteq \{x{:}xR_ix \text{ for all } i\}$, then $\{x{:}aR^{\#*}x\} \subseteq \{x{:}xR_i{}^{\#}x \text{ for all } i\}$. This is clear, since $\{x{:}aR^{\#*}x\} \subseteq \{x{:}aR^{*}x) \cap W^{\#}$, and $(x{:}xR_i{}^{\#}x \text{ for all } i) = \{x{:}xR_ix \text{ for all } i\} \cap W^{\#}$. Second, because of the fact that players know all previous moves at the beginning of each stage, they can make their strategy choices conditional on whether a subgame is reached. (More precisely, for any player i and pair of strategies s and s′ for i, that are compatible with $\Gamma^{\#}$ being reached, there is a strategy equivalent to this: s if $\Gamma^{\#}$ is reached, s′ if not.) This implies that for any world w, player i and subgame such that it is compatible with i′s beliefs that that subgame be reached, a strategy will be rational for i only if the strategy determined for the subgame is rational, conditional on the hypothesis that

the subgame is reached. This ensures that rationality is preserved in all worlds when the model is modified. So $c^{\#}$ is strongly rationalizable in $\Gamma^{\#}$.

An analogous result about strong *perfect* rationalizability can be shown by essentially the same argument.

One further observation before turning to the backward induction theorem itself: for any game Γ, if profile c is compatible with common belief in (the actual world of) an SR model for Γ, then c itself is strongly rationalizable. It is obvious that if $S(x) = c$ and aR^*x, then the same model, with x rather than a as the actual world will be an SR model if the original model was.

Now the backward induction theorem:

Any strongly rationalizable strategy profile in a finite iterated prisoners' dilemma is one in which both players defect every time.

The proof is by induction on the size of the game. For the base case – the one shot PD – it is obvious that the theorem holds, since only defection is rational. Now assume that the theorem holds for games of length k. Let Γ be a game of length $k + 1$, and Γ^- be the corresponding iterated PD of length k. Let M be any SR model of Γ, and let c be any strategy profile that is compatible with common belief (that is, c is any profile for which there exists an x such that $S(x) = c$, and aR^*x). By the observation just made, c is strongly rationalizable, so by the multi-stage game lemma, c^- (the profile for Γ^- determined by c) is strongly rationalizable in Γ^-. But then by hypothesis of induction, c^- is a profile in which both players defect every time. So c (in game Γ) is a profile in which both players defect every time after the first move. But c is any profile compatible with common belief in the actual world of the model, so it follows that in the model M, it is common belief that both players will choose strategies that result in defection every time after the first move. Given these beliefs, any strategy for either player that begins with the cooperative move is strictly dominated, relative to that player's beliefs. So since the players are both rational, it follows that they choose a strategy that begins with defection, and so one that results in defection on every move.

Our theorem could obviously be generalized to cover some other games that have been prominent in discussions of backward induction such as the centipede game and (for strong *perfect* rationalizability) the chain store game. But it is not true, even in perfect information games, that the strong or strong and perfect rationalizability conditions are always sufficient to support backward induction reasoning. Recall the perfect information, pure coordination game discussed above in which Alice and Bert failed to coordinate on the backward induction equilibrium, even though the conditions for strong perfect rationalizability were satisfied. In that example, the strategy profile played was a perfect, but not subgame perfect, equilibrium. One can show in general that in perfect information games, all and only Nash equilibrium strategy profiles are strongly rationalizable (see Stalnaker (1994) for the proof).

As I noted at the end of the last section, it can be shown that the set of strongly and perfectly rationalizable strategy profiles is characterized also by the class of models in which there is common knowledge (in the defeasibility sense) of perfect rationality. So we can drop the strong idealizing assumption that there

is no error, and still get the conclusion that if there is common knowledge (in the defeasibility sense) of perfect rationality, then players will choose strategies that result in defection every time.

Pettit and Sugden, in their discussion of the paradox of backward induction, grant that the argument is valid when it is common knowledge rather than common belief that is assumed (though they don't say why they think this, or what they are assuming about knowledge). But they suggest that there is nothing surprising or paradoxical about this, since the assumption of common *knowledge* of rationality is incompatible with the possibility of rational deliberation, and so is too strong to be interesting. Since knowledge logically implies truth, they argue, the argument shows that 'as a matter of logical necessity, both players *must* defect and presumably therefore that they know they must defect' (Pettit and Sugden 1989). But I think this remark rests on a confusion of epistemic with causal possibilities. There is no reason why I cannot both *know* that something is true, and also entertain the counterfactual possibility that it is false. It is of course inconsistent to suppose, counterfactually or otherwise, the conjunction of the claim that ϕ is false with the claim that I know that ϕ is true, but it is not inconsistent for me, knowing (in the actual world) that ϕ is true, to suppose, counterfactually, that ϕ is false. As Pettit and Sugden say, the connection between knowledge and truth is a matter of logical necessity, but that does not mean that if I know that I will defect, I therefore *must* defect, 'as a matter of logical necessity'. One might as well argue that lifelong bachelors are powerless to marry, since it is a matter of logical necessity that lifelong bachelors never marry.

The semantic connection between knowledge and truth is not, in any case, what is doing the work in this version of the backward induction argument: it is rather the assumption that the players believe in common that neither of them is in error about anything. We could drop the assumption that the players beliefs are all actually true, assuming not common knowledge of rationality, but only common belief in rationality and common belief that no one is in error about anything. This will suffice to validate the induction argument.

Notice that the common belief that there will not, in fact, be any surprises, does not imply the belief that there couldn't be any surprises. Alice might think as follows: 'Bert expects me to defect, and I will defect, but I could cooperate, and if I did, he would be surprised. Furthermore, I expect him to defect, but he could cooperate, and if he did, I would be surprised'. If these 'could's were epistemic or subjective, expressing uncertainty, then this soliloquy would make no sense, but it is unproblematic if they are counterfactual 'could's used to express Alice's beliefs about her and Bert's capacities. A rational person may know that she will not exercise certain of her options, since she may believe that it is not in her interest to do so.

It is neither legitimate nor required for the success of the backward induction argument to draw conclusions about what the players would believe or do under counterfactual conditions. In fact, consider the following 'tat for tit' strategy: defect on the first move, then on all subsequent moves, do what the other player did on the previous move, until the last move; defect unconditionally on the last move. Our backward induction argument does not exclude the possibility that the players

should each adopt, in the actual world, this strategy, since this pair of strategies results in defection every time. This pair is indeed compatible with the conditions for strong and perfect rationalizability. Of course unless each player assigned a very low probability to the hypothesis that this was the other player's strategy, it would not be rational for him to adopt it, but he need not rule it out. Thus Pettit and Sugden are wrong when they say that the backward induction argument can work only if it is assumed that each player would maintain the beliefs necessary for common belief in rationality 'regardless of what the other does' (Pettit and Sugden 1989, p. 178). All that is required is the belief that the beliefs necessary for common knowledge of rationality will, in fact, be maintained, given what the players in fact plan to do. And this requirement need not be *assumed*: it is a consequence of what is assumed.

Conclusion

The aim in constructing this model theory was to get a framework in which to sharpen and clarify the concepts used both by rational agents in their deliberative and strategic reasoning and by theorists in their attempts to describe, predict and explain the behavior of such agents. The intention was, first, to get a framework that is rich in expressive resources, but weak in the claims that are presupposed or implicit in the theory, so that various hypotheses about the epistemic states and behavior of agents can be stated clearly and compared. Second, the intention was to have a framework in which concepts can be analyzed into their basic components, which can then be considered and clarified in isolation before being combined with each other. We want to be able to consider, for example, the logic of belief, individual utility maximization, belief revision, and causal-counterfactual structure separately, and then put them together to see how the separate components interact. The framework is designed to be extended, both by considering further specific substantive assumptions, for example, about the beliefs and belief revision policies of players, and by adding to the descriptive resources of the model theory additional structure that might be relevant to strategic reasoning or its evaluation, for example temporal structure for the representation of dynamic games, and resources for more explicit representation of counterfactual propositions. To illustrate some of the fruits of this approach we have stated some theorems that provide model theoretic characterizations of some solution concepts, and have looked closely at one familiar form of reasoning – backward induction – and at some conditions that are sufficient to validate this form of reasoning in certain games, and at conditions that are not sufficient. The focus has been on the concepts involved in two kinds of counterfactual reasoning whose interaction is essential to deliberation in strategic contexts, and to the evaluation of the decisions that result from such deliberation: reasoning about what the consequences would be of actions that are alternatives to the action chosen, and reasoning about how one would revise one's beliefs if one were to receive information that one expects not to receive. We can get clear about why

people do what they do, and about what they ought to do, only by getting clear about the relevance of what they could have done, and might have learned, but did not.[21]

References

Adams, E. (1970). Subjunctive and indicative conditionals. *Foundations of Language, 6*, 89–94.
Alchourón, C., & Makinson, D. (1982). The logic of theory change: Contraction functions and their associated revision functions. *Theoria, 48*, 14–37.
Alchourón, C., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic, 50*, 510–530.
Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics, 4*, 1236–1239.
Bacharach, M. (1985). Some extensions of a claim of Aumann in an axiomatic model of knowledge. *Journal of Economic Theory, 37*, 167–190.
Bernheim, B. (1984). Rationalizable strategic behavior. *Econometrica, 52*, 1007–1028.
Blume, L., Brandenburger, A., & Dekel, E. (1991a). Lexicographic probabilities and choice under uncertainty. *Econometrica, 59*, 61–79.
Blume, L., Brandenburger, A., & Dekel, E. (1991b). Lexicographic probabilities and equilibrium refinements. *Econometrica, 59*, 81–98.
Dekel, E., & Fudenberg, D. (1990). Rational behavior with payoff uncertainty. *Journal of Economic Theory, 52*, 243–267.
Fudenberg, D., & Tirole, J. (1992). *Game theory*. Cambridge, MA: MIT Press.
Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge, MA: MIT Press.
Gibbard, A., & Harper, W. (1981). Counterfactuals and two kinds of expected utility. In C. Hooker et al. (Eds.), *Foundations and applications of decision theory*. Dordrecht/Boston: Reidel.
Grove, A. (1988). Two modelings for theory change. *Journal of Philosophical Logic, 17*, 157–170.
Harper, W. (1975). Rational belief change, popper functions and counterfactuals. *Synthese, 30*, 221–262.
Lewis, D. (1980). Causal decision theory. *Australasian Journal of Philosophy, 59*, 5–30.
Makinson, D. (1985). How to give it up: A survey of some formal aspects of the logic of theory change. *Synthese, 62*, 347–363.
Mongin, P. (1994). The logic of belief change and nonadditive probability. In D. Prawitz & D. Westerstahl (Eds.), *Logic and philosophy of science in Uppsala*. Dordrecht: Kluwer.
Pappas, G., & Swain, M. (1978). *Essays on knowledge and justification*. Ithaca: Cornell University Press.
Pearce, G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica, 52*, 1029–1050.
Pettit, P., & Sugden, R. (1989). The backward induction paradox. *Journal of Philosophy, 86*, 169–182.
Skyrms, B. (1982). Causal decision theory. *Journal of Philosophy, 79*, 695–711.
Skyrms, B. (1992). *The dynamics of rational deliberation*. Cambridge, MA: Harvard University Press.
Spohn, W. (1987). Ordinal conditional functions: A dynamic theory of epistemic states. In W. Harper & B. Skyrms (Eds.), *Causation in decision, belief change and statistics* (Vol. 2, pp. 105–134). Dordrecht: Reidel.
Stalnaker, R. (1994). On the evaluation of solution concepts. *Theory and Decision, 37*, 49–73.

[21]I would like to thank Pierpaolo Battigalli, Yannis Delmas, Drew Fudenberg, Philippe Mongin, Hyun Song Shin, Brian Skyrms, and an anonymous referee for helpful comments on several earlier versions of this paper.

Chapter 43
Substantive Rationality and Backward Induction

Joseph Y. Halpern

Starting with the work of Bicchieri (1988, 1989), Binmore (1987), and Reny (1992), there has been intense scrutiny of the assumption of common knowledge of rationality, the use of counterfactual reasoning in games, and the role of common knowledge and counterfactuals in the arguments for backward induction in games of perfect information. Startlingly different conclusions were reached by different authors.

These differences were clearly brought out during a 2.5 h round table discussion on "Common knowledge of rationality and the backward induction solution for games of perfect information" at the 1998 TARK (Theoretical Aspects of Rationality and Knowledge) conference. During the discussion, Robert Aumann and Robert Stalnaker stated the following theorems:

Aumann's Theorem (Informal version). *Common knowledge of substantive rationality implies the backwards induction solution in games of perfect information.*

Stalnaker's Theorem (Informal version). *Common knowledge of substantive rationality does not imply the backwards induction solution in games of perfect information.*

The discussion during the round table was lively, but focused on more philosophical, high-level issues. My goal in this short note is to explain the technical differences between the framework of Aumann (1995) and Stalnaker (1996) that lead to the different results, and to show what changes need to be made to Aumann's

Supported in part by NSF under grant IRI-96-25901.

J.Y. Halpern (✉)
Computer Science Department, Cornell University, Ithaca, NY 14853, USA
e-mail: halpern@cs.cornell.edu

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_43

framework to get Stalnaker's result.[1] I believe that the points that I make here are well known to some (and certainly were made informally during the discussion). Indeed, many of the key conceptual points I make already appear in Stalnaker's discussion of Aumann's result in Stalnaker (1998, pp. 45–50). However, since Stalnaker uses belief rather than knowledge and must deal with the complications of having probability, it is not so easy to directly compare his results in Stalnaker (1998) with Aumann's. I hope that the simpler model I present here will facilitate a careful comparison of the differences between Aumann's and Stalnaker's results and thus clarify a few issues, putting the debate on a more rational footing.

There are three terms in the theorems that need clarification:

- (common) knowledge
- rationality
- *substantive* rationality

I claim that Stalnaker's result can be obtained using exactly the same definition of (common) knowledge and rationality as the one Aumann (1995) used. The definition of knowledge is the standard one, given in terms of partitions. (I stress this point because Stalnaker (1996) has argued that probability-1 belief is more appropriate than knowledge when considering games.) The definition of rationality is that a player who uses strategy s is rational at vertex v if there is no other strategy that he knows will give him a better payoff, conditional on being at vertex v. Both Aumann and Stalnaker give substantive rationality the same reading: "rationality at all vertices v in the game tree". They further agree that this involves a counterfactual statement: "for all vertices v, if the player were to reach vertex v, then the player would be rational at vertex v". The key difference between Aumann and Stalnaker lies in how they interpret this counterfactual. In the rest of this note, I try to make this difference more precise.

The Details

I start by considering Aumann's model. Fix a game Γ of perfect information for n players. As usual, we think of Γ as a tree. Because Γ is a game of perfect information, the players always know which vertex in the tree describes the current situation in the game. The nonleaf vertices in Γ are partitioned into n sets, $G_1, \ldots, G_n$, one for each player. The vertices in G_i are said to belong to i; these are the ones where player i must move. A *model of* Γ is a tuple $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s})$, where Ω is a set of states of the world, $\mathcal{K}_1, \ldots, \mathcal{K}_n$ are partitions, one for each player $i = 1, \ldots, n$ ($\mathcal{K}_i$ is i's information partition), and $\mathbf{s}$ maps each world $\omega \in \Omega$

[1] The model that I use to prove Stalnaker's result is a variant of the model that Stalnaker (1996) used, designed to be as similar as possible to Aumann's model so as to bring out the key differences. This, I believe, is essentially the model that Stalnaker had in mind at the round table.

to a strategy profile $\mathbf{s}(\omega) = (s_1, \ldots, s_n)$; s_i is i's strategy in game Γ at state ω. As usual, a strategy for i in Γ is just a mapping from i's vertices in Γ to actions. I write $\mathbf{s}_i(\omega)$ for s_i.

Let $\mathcal{K}_i(\omega)$ denote the cell in partition $\mathcal{K}_i$ that includes ω. Define the operator K_i on events as usual:

$$K_i(E) = \{\omega : \mathcal{K}_i(\omega) \subseteq E\}.$$

$K_i(E)$ is the event that i knows E. Let $A(E) = K_1(E) \cap \ldots \cap K_n(E)$. $A(E)$ is the event that everyone (*all* the players) know E. Let

$$CK(E) = A(E) \cap A(A(E)) \cap A(A(A(E))) \cap \ldots$$

$CK(E)$ is the event that E is common knowledge.

Aumann and Stalnaker (and everyone else who has written on this subject that I am aware of) assume that the players know their strategies. Formally, that means that if $\omega' \in \mathcal{K}_i(\omega)$, then $\mathbf{s}_i(\omega) = \mathbf{s}_i(\omega')$; that is, i uses the same strategy at all the states in a cell of $\mathcal{K}_i$.

Next we need to define rationality. Note that a strategy profile s and vertex v uniquely determine a path in Γ that would be followed if s were played starting at v. Let $h_i^v(s)$ denote i's payoff if this path is followed. Informally, i is rational at vertex v if there is no strategy that i could have used that i knows would net him a higher payoff than the strategy he actually uses. More precisely, *i is rational at vertex v in ω* if, for all strategies $s^i \neq \mathbf{s}_i(\omega)$, $h_i^v(\mathbf{s}(\omega')) \geq h_i^v(\mathbf{s}_{-i}(\omega'), s^i))$ for some $\omega' \in \mathcal{K}_i(\omega)$. That is, i cannot do better by using s^i than $\mathbf{s}_i(\omega)$ against all the strategy profiles of the other players that he considers possible at ω. This is a weak notion of rationality (which is certainly satisfied by expected utility maximization). By taking such a weak notion, Aumann's Theorem becomes stronger. As will be clear from the example, Stalnaker's Theorem holds even if we strengthen the requirements of rationality (to require strict inequality, for example).

Aumann then defines *substantive rationality* to mean rationality at all vertices in the game tree. That is, i is substantively rational in state ω if i is rational at vertex v in ω for every vertex $v \in G_i$. For future reference, I call Aumann's notion of substantive rationality A-rationality.

Using these definitions, Aumann can and does prove his theorem (using a straightforward backward induction argument).

Stalnaker's definition of substantive rationality is different from Aumann's although, as I indicated above, he is trying to capture the same intuition. His definition tries to enforce the intuition that, for every vertex $v \in G_i$, *if* i were to actually reach v, then what he would do in that case would be rational. The key point is that, according to Stalnaker's definition, in order to evaluate, at state ω, whether i is being rational at vertex v by performing the action dictated by his strategy at ω, we must consider i's beliefs in the state "closest" to ω according to i where v is actually reached.

To formalize this, we must add one more component to Aumann's model: for each player i, we must have a selection function f mapping states and i's

vertices to states. An *extended model of* Γ is a tuple $(\Omega, \mathscr{K}_1, \ldots, \mathscr{K}_n, \mathbf{s}, f)$, where $(\Omega, \mathscr{K}_1, \ldots, \mathscr{K}_n, \mathbf{s})$ is a model of Γ, $f : \Omega \times G \to \Omega$, and $G = \cup_{i=1}^n G_i$. Intuitively, if $f(\omega, v) = \omega'$, then state ω' is the state closest to ω where vertex v is reached.[2]

Given this intuition, we may want to assume that f satisfies some constraints such as the following:

F1. v is reached in $f(\omega, v)$ (that is, v is on the path determined by $\mathbf{s}(f(\omega, v))$).
F2. If v is reached in ω, then $f(\omega, v) = \omega$.
F3. $\mathbf{s}(f(\omega, v))$ and $\mathbf{s}(\omega)$ agree on the subtree of Γ below v.

F1 guarantees that v is actually reached in $f(\omega, v)$, while F2 says that if v is actually reached in ω, then ω is the closest state to itself where v is reached. F3 is intended to capture the intuitive meaning of a strategy. If, according to $\mathbf{s}_i(\omega)$, player i performs action a at vertex v, it seems reasonable to expect that at the closest world where v is actually reached, player i does in fact perform a. This follows from F3. However, F3 says more than this. It says that at all the vertices below v, all the players also perform the actions dictated by $\mathbf{s}(\omega)$. This extra requirement arguably makes F3 too strong. However, as I shall show, Stalnaker's Theorem continues to hold even with this strong assumption.[3]

According to Stalnaker, i is substantively rational in state ω if i is rational at vertex v in $f(\omega, v)$ for every vertex $v \in G_i$. Let us call this notion *S-rationality*. Thus, the crucial (and, in fact, only) difference between Aumann's approach and Stalnaker's approach is that A-rationality requires i to be rational at vertex v in ω and S-rationality requires i to be rational at vertex v in $f(\omega, v)$. Notice that, by F3, if it is i's move at vertex v, then i makes the same move at both ω and $f(\omega, v)$. However, this move may be rational at $f(\omega, v)$ and irrational at ω, since, in general, i's beliefs are different at ω and $f(\omega, v)$.

The difference between A-rationality and S-rationality can perhaps be best understood by considering the game described in Fig. 43.1, which is also considered by Stalnaker (1998) and is a variant of a game introduced by Aumann (1995).

Consider the following five strategy profiles:

- s^1 is the strategy profile (da, d) (i.e., Ann goes down at v_1 and across at v_3 and Bob goes down at v_2);
- s^2 is the strategy profile (aa, d);

[2] Again, I should stress that this is not exactly the model that Stalnaker uses in (1996), but it suffices for my purposes. I remark that in Halpern (1999), I use selection functions indexed by the players, so that player 1 may have a different selection function than player 2. I do not need this greater generality here, so I consider the simpler model where all players use the same selection function.

[3] There are certainly other reasonable properties we could require of the selection function. For example, we might want to require that if v is reached in some state in $\mathscr{K}_i(\omega)$, then $f(\omega, v) \in \mathscr{K}_i(\omega)$. I believe that it is worth trying to characterize the properties we expect the selection function should have, but this issue would take us too far afield here. (See Stalnaker RC, 1999, Counterfactual propositions in games, Unpublished manuscript, for further discussion of this point.) Note that F1–F3 are properties that seem reasonable for arbitrary games, not just games of perfect information.

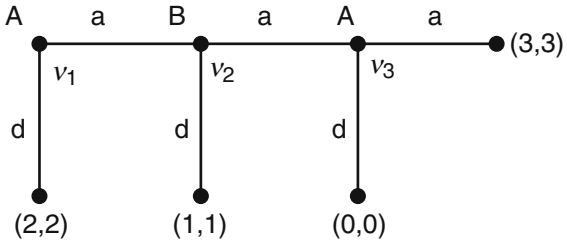


Fig. 43.1 A variant of Aumann's Ann-Bob game

- s^3 is the strategy profile (ad, d);
- s^4 is the strategy profile (aa, a);
- s^5 is the strategy profile (ad, a).

Note that s^4 is the backward induction solution.

Now consider the extended model $(\omega, \mathcal{K}_{Ann}, K_{Bob}, \mathbf{s}, f)$ of this game, where

- $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$;
- $\mathcal{K}_{Ann}(\omega_i) = \{\omega_i\}$, for $i = 1, \ldots, 5$;
- $\mathcal{K}_{Bob}(\omega_i) = \{\omega_i\}$ for $i = 1, 4, 5$; $\mathcal{K}_{Bob}(\omega_2) = \mathcal{K}_{Bob}(\omega_3) = \{\omega_2, \omega_3\}$;
- $\mathbf{s}(\omega_i) = s^i$, for $i = 1, \ldots, 5$;
- f is the unique selection function satisfying F1–F3 (so that $f(\omega_i, v_1) = \omega_i$ for all i; $f(\omega_1, v_2) = \omega_2$ and $f(\omega_i, v_2) = \omega_i$ for $i = 2, \ldots, 5$; $f(\omega_1, v_3) = f(\omega_2, v_2) = \omega_4$, $f(\omega_3, v_3) = \omega_5$, and $f(\omega_i, v_3) = \omega_i$ for $i = 4, 5$).

It is easy to check that, at ω_1, it is common knowledge that strategy profile s^1 is being used. It is also common knowledge at ω_1 that, if vertex v_2 were reached, Bob would play down.

In this extended model, clearly Bob is not rational at vertex v_2 in ω_1, since he plays down. This means that we do not have A-rationality at ω_1 (and, a fortiori, we do not have common knowledge of A-rationality at ω_1). On the other hand, Bob is rational at vertex v_2 in ω_2, since Bob considers it possible that Ann may go down at v_3 (since $\mathcal{K}_{Bob}(\omega_2) = \{\omega_2, \omega_3\}$). Similarly, Ann is rational at vertex v_3 in ω_4. Since $f(\omega_1, v_2) = \omega_2$ and $f(\omega_1, v_3) = \omega_4$, it follows that we have S-rationality at ω_1, and hence common knowledge of S-rationality at ω_1.

This example is an instance of Stalnaker's Theorem: we have common knowledge of substantive rationality in the sense of S-rationality at ω_1, yet the backward induction solution is not played at ω_1. Nevertheless, it does not contradict Aumann's Theorem, since we do not have common knowledge of A-rationality.

With this machinery, we can now state Aumann's Theorem and Stalnaker's Theorem more formally. Let *S-RAT* consist of all states where all the players are S-rational; let *A-RAT* consist of all states where all the players are A-rational; let *BI* consist of all states where the backward induction solution is played.

Aumann's Theorem. *If* Γ *is a nondegenerate*[4] *game of perfect information, then in all models of* Γ*, we have* $CK(A\text{-}RAT) \subseteq BI$.

Stalnaker's Theorem. *There exists a nondegenerate game* Γ *of perfect information and an extended model of* Γ *in which the selection function satisfies F1–F3 such that* $CK(S\text{-}RAT) \not\subseteq BI$.

Note that, in an extended model of the Ann-Bob game, it is consistent for Ann to say "Although it is common knowledge that I would play across if v_3 were reached, if I were to play across at v_1, Bob would consider it possible that I would play down at v_3." This is not possible in Aumann's framework because, without selection functions, Aumann has no way of allowing the players to revise their beliefs. (This point is essentially made by Samet (1996).) In the definition of A-rationality, for any vertex v, player i's beliefs in state ω about the possible strategies player j may be using if vertex v is reached are the same (and are determined by $\mathcal{K}_i(\omega)$). It is crucial for Aumann's result (and, I believe, a weakness in his model) that players do not (and cannot) revise their beliefs about other players' strategies when doing such hypothetical reasoning.

It is not hard to place a condition on selection functions that guarantees that players' beliefs about other players' strategies do not change when they are doing hypothetical reasoning.

F4. For all players i and vertices v, if $\omega' \in \mathcal{K}_i(f(\omega, v))$ then there exists a state $\omega'' \in \mathcal{K}_i(\omega)$ such that $\mathbf{s}(\omega')$ and $\mathbf{s}(\omega'')$ agree on the subtree of Γ below v.[5]

Combined with F1–F3, F4 this gives us the properties we want. In fact, we have the following result, similar in spirit to a result proved by Stalnaker (1998, p. 43).

Theorem 1. *If* Γ *is a nondegenerate game of perfect information, then for every extended model of* Γ *in which the selection function satisfies F1–F4, we have* $CK(S\text{-}RAT) \subseteq BI$. *Moreover, there is an extended model of* Γ *in which the selection function satisfies F1–F4.*

Proof. For the first half of the theorem, suppose $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$ is an extended model of Γ in which the selection function satisfies F1–F4. Further suppose that $\omega \in CK(S\text{-}RAT)$. We want to show $\omega \in BI$. The proof basically mimics Aumann's proof of his theorem, since F1–F4 essentially gives us his framework.

I first recall a standard characterization of common knowledge. Define the notion of ω'' *being reachable from* ω' *in* k *steps* inductively: ω'' *is reachable from* ω' *in 1 step* iff $\omega'' \in \mathcal{K}_i(\omega')$ for some $i \in \{1, \ldots, n\}$; ω'' *is reachable from* ω' *in* $k + 1$

[4] A game is nondegenerate if the payoffs are different at all the leaves.

[5] Actually, F4 says that player i considers at least as many strategies possible at ω as at $f(\omega, v)$. To capture the fact that player i's beliefs about other players' possible strategies do not change, we would need the opposite direction of F4 as well: if $\omega' \in \mathcal{K}_i(\omega)$ then there exists a state $\omega'' \in \mathcal{K}_i(f(\omega, v))$ such that $\mathbf{s}(\omega')$ and $\mathbf{s}(\omega'')$ agree on the subtree of Γ below v. I do not impose this requirement here simply because it turns out to be unnecessary for Aumann's Theorem.

steps iff there exists a state ω''' that is reachable from ω' in 1 step such that ω'' is reachable from ω''' in k steps. We say that ω'' *is reachable from* ω' if ω'' is reachable from ω' in k steps for some k. It is well known (Aumann 1976) that $\omega' \in CK(E)$ iff $\omega'' \in E$ for all ω'' reachable from ω'.

I show by induction on k that for all states ω' reachable from ω, if v is a vertex which is at height k in the game tree (i.e., k moves away from a leaf), the move dictated by the backward induction solution (for the subgame of Γ rooted at v) is played at v in state ω'.

For the base case, suppose v is at height 1 and ω' is reachable from ω. Since $\omega \in CK(S\text{-}RAT)$, we must have $\omega' \in S\text{-}RAT$. Suppose player i moves at ω'. Since $\omega' \in S\text{-}RAT$, player i must make the move dictated by the backwards induction solution at $f(\omega', v)$. By F3, he must do so at ω' as well.

For the inductive step, suppose that v is at height $k+1$, player i moves at v, and ω' is reachable from ω. Suppose, by way of contradiction, that a is the action indicated by the backward induction solution at v but $\mathbf{s}_i(\omega')(v) = a' \neq a$. Note that by the induction hypothesis, at every vertex below v, all the players play according to the backward induction solution in state ω'. Since $\omega' \in S\text{-}RAT$, we must have that i is rational at v in $f(\omega', v)$. By F3, it follows that i plays a' at vertex v in $f(\omega', v)$ and at every vertex below v, the players play according to the backward induction solution. Thus, there must be a state $\omega'' \in \mathcal{K}_i(f(\omega', v))$ such that by using $\mathbf{s}_i(f(\omega', v))$, player i does at least as well in ω'' as by using the backward induction strategy starting from v. By F4, there must exist some $\omega''' \in \mathcal{K}_i(\omega')$ such that $\mathbf{s}(\omega'')$ and $\mathbf{s}(\omega''')$ agree on the subtree of Γ below v. Since ω''' is reachable from ω, by the induction hypothesis, all players use the backward induction solution at vertices below v. By F3, this is true at ω'' as well. However, this means that player i does better at ω'' playing a at v than a', giving us the desired contradiction.

For the second half, given a nondegenerate game Γ of perfect information, let s be the strategy where, at each vertex v, the players play the move dictated the backward induction solution in the game defined by the subtree below v. For each vertex v, let s_v be the strategy where the players play the actions required to reach vertex v, and then below v, they play according to s. Note that if v is reached by s, then $s_v = s$. In particular, if r is the root of the tree, then $s_r = s$. Consider the extended model $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$ where $\Omega = \{\omega_v : v \text{ is a vertex of } \Gamma\}$, $\mathcal{K}_i(\omega_v) = \{\omega_v\}$, $\mathbf{s}(\omega_v) = s_v$, and $f(\omega, v)$ is ω if v is reached by $\mathbf{s}(\omega)$ and ω_v otherwise. I leave it to check that this gives us an extended model where the selection function satisfies F1–4. ■

Remarks. 1. As mentioned earlier, Theorem 1 is similar in spirit to the theorem proved by Stalnaker (1998, p. 43). To distinguish this latter theorem from what I have been calling "Stalnaker's Theorem" in this paper, I call the latter theorem "Stalnaker's probabilistic theorem", since it applies to his formal model with probabilities. To state Stalnaker's probabilistic theorem precisely, I need to explain some differences between my model and his. First, as I mentioned earlier, Stalnaker considers belief rather than knowledge, where belief is essentially "with probability 1". He defines a probabilistic analogue of extended models and

a notion of *perfect rationality* in such models which is an analogue of substantive rationality. In addition, Stalnaker considers an epistemic independence condition on the players' beliefs. This condition says that a player treats information about disjoint sets of players as epistemically independent. This means, in particular, that if v was the result of player i's move, then a player's belief about the rationality of players other than i will be the same in ω and $f(\omega, v)$.

F4 plays essentially the same role as this independence condition, although formally it is somewhat different. F4 says that the set of strategies a player considers possible at $f(\omega, v)$ is a subset of that he considers possible in ω. The stronger version (discussed in Footnote 5) says that he considers possible the same set of strategies in ω and $f(\omega, v)$. The stronger version can be viewed as saying that a player's beliefs about the strategies that will be used below v (including his own) is epistemically independent of the move made.

In any case, with this background, I can state Stalnaker's probabilistic theorem. It also gives a condition sufficient to ensure that the backward induction solution is played. It says that if Γ is a nondegenerate game of perfect information in which each player moves at most once and M is an extended model (in Stalnaker's sense) of Γ where it is common belief that all players are perfectly rational and players treat information about disjoint sets of players as epistemically independent, then the backwards induction solution is played in M.

The restriction in Stalnaker's Theorem to games where each player moves at most once is not that serious. Given an arbitrary game Γ, we can consider the *agent-normal form* Γ' of Γ (Selten 1975), by associating a different "self" of player i with each vertex of Γ where player i moves. We view these different selves as different agents. Given an extended model M of Γ, let M' be the corresponding extended model of Γ' where, to each world ω in M, there is a corresponding world ω' in M' such that each "self" of player i is endowed with the same beliefs at world ω' as player i has at ω. M' satisfies Stalnaker's epistemic independence assumption; since in Γ' each player makes at most one move, Stalnaker's probabilistic theorem applies. Note that the epistemic independence assumption now says player i's early moves do not affect other players' beliefs about whether player i will player rationally at later in the game (since the later vertices are under the control of different agents—different selves of agent i). Thus, it should be clear that epistemic independence plays essentially the same role in Stalnaker's result as F4 does in Theorem 1.

2. Theorem 1 and the earlier discussion suggests that one possible culprit for the confusion in the literature regarding what is required to force the backwards induction solution in games of perfect information is the notion of a *strategy*. Exactly what should it mean to say that Ann's strategy at a state ω is s? For example, consider the game in Fig. 43.1. According to strategy s_A^1, Ann plays across at vertex v_3. But v_3 is a vertex that cannot be reached if Ann uses s_A^1, since according to this strategy, Ann plays down at v_1. The standard reading of $s_A^1(v_3) = a$ is that "*if* v_3 is reached, then Ann plays across". But this reading leaves a number of questions unanswered. How Ann plays (if she is rational) depends on her beliefs. Should we read this as "no matter what Ann's beliefs are,

if v_3 is reached, Ann will play a"? Or perhaps it should be "given her current beliefs (regarding, for example, what move Bob will make), if v_3 is reached, Ann will play a". Or perhaps it should be "in the state 'closest' to the current state where v_3 is actually reached, Ann plays a".

I have taken the last reading here (where 'closest' is defined by the selection function); assumption F3 essentially forces it to be equivalent to the second reading.

However, without F4, this equivalence is not maintained with regard to Bob's beliefs. That is, consider the following two statements:

- Bob currently believes that, given Ann's current beliefs, Ann will play a if v_3 is reached;
- in the state closest to the current state where v_3 is reached, Bob believes that Ann plays a at v_3.

The first statement considers Bob's beliefs at the current state; the second considers Bob's beliefs at a different state. Without F4, these beliefs might be quite different. It is this possible difference that leads to Stalnaker's Theorem.

3. Strategies themselves clearly involve counterfactual reasoning. If we take strategies as primitive objects (as both Aumann and Stalnaker do, and as I have done for consistency), we have two sources of counterfactuals in extended models: selection functions and strategies. Stalnaker (1996, p. 135) has argued that "To clarify the causal and epistemic concepts that interact in strategic reasoning, it is useful to break them down into their component parts." This suggests that it would be useful to have a model where strategy is *not* a primitive, but rather is defined in terms of counterfactuals. This is precisely what Samet (1996) does.[6]

   Not surprisingly, in Samet's framework, Aumann's Theorem does not hold without further assumptions. Samet shows that what he calls a *common hypothesis* of rationality implies the backward induction solution in nondegenerate games of perfect information. Although there are a number of technical differences in the setup, this result is very much in the spirit of Theorem 1.

Acknowledgements I'd like to thank Robert Stalnaker for his many useful comments and criticisms of this paper.

References

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics, 4*(6), 1236–1239.

Aumann, R. J. (1995). Backwards induction and common knowledge of rationality. *Games and Economic Behavior, 8*, 6–19.

[6]Samet does not use selection functions to capture counterfactual reasoning, but *hypothesis transformations*, which map cells (in the information partition) to cells. However, as I have shown (Halpern 1999), we can capture what Samet is trying to do by using selection functions.

Bicchieri, C. (1988). Strategic behavior and counterfactuals. *Synthese, 76*, 135–169.
Bicchieri, C. (1989). Self refuting theories of strategic interaction: A paradox of common knowledge. *Erkenntnis, 30*, 69–85.
Binmore, K. (1987). Modeling rational players I. *Economics and Philosophy, 3*, 179–214. Part II appeared ibid., *4*, 9–55.
Halpern, J. Y. (1999). Hypothetical knowledge and counterfactual reasoning. *International Journal of Game Theory, 28*(3), 315–330.
Reny, P. (1992). Rationality in extensive form games. *Journal of Economic Perspectives, 6*, 103–118.
Samet, D. (1996). Hypothetical knowledge and games with perfect information. *Games and Economic Behavior, 17*, 230–251.
Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory, 4*, 25–55.
Stalnaker, R. (1998). Belief revision in games: Forward and backward induction. *Mathematical Social Sciences, 36*, 31–56.
Stalnaker, R. C. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy, 12*, 133–163.

Index

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2