THEORY AND DECISION LIBRARY

General Editor: Julian Nida-Rümelin (*Universität München*)

Series A: Philosophy and Methodology of the Social Sciences

Series B: Mathematical and Statistical Methods

Series C: Game Theory, Mathematical Programming and Operations Research

SERIES A: PHILOSOPHY AND METHODOLOGY
OF THE SOCIAL SCIENCES

VOLUME 42

Assistant Editor: Martin Rechenauer (*Universität München*)

Editorial Board: Raymond Boudon (*Paris*), Mario Bunge (*Montréal*), Isaac Levi (*New York*), Richard V. Mattessich (*Vancouver*), Bertrand Munier (*Cachan*), Amartya K. Sen (*Cambridge*), Brian Skyrms (*Irvine*), Wolfgang Spohn (*Konstanz*)

Scope: This series deals with the foundations, the general methodology and the criteria, goals and purpose of the social sciences. The emphasis in the Series A will be on well-argued, thoroughly analytical rather than advanced mathematical treatments. In this context, particular attention will be paid to game and decision theory and general philosophical topics from mathematics, psychology and economics, such as game theory, voting and welfare theory, with applications to political science, sociology, law and ethics.

For other titles published in this series, go to
http://www.springer.com/series/6616

PREFERENCE CHANGE
Approaches from Philosophy, Economics and Psychology

Edited by

TILL GRÜNE-YANOFF and SVEN OVE HANSSON
Collegium of Advanced Studies, Helsinki
Royal Institute of Technology, Stockholm

Springer

Editors
Till Grüne-Yanoff
Helsinki Collegium of Advanced Studies
Fabiankatu 24
00014 University of Helsinki
Finland
till.grune@helsinki.fi

Sven Ove Hansson
Royal Institute of Technology
Division of Philosophy
SE-100 44 Stockholm
Sweden
soh@kth.se

ISBN 978-90-481-2592-0 e-ISBN 978-90-481-2593-7
DOI 10.1007/978-90-481-2593-7
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009926166

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Changing preferences is a phenomenon often invoked but rarely properly accounted for. Throughout the history of the social sciences, researchers have come against the possibility that their subjects' preferences were affected by the phenomena to be explained or by other factors not taken into account in the explanation. Sporadically, attempts have been made to systematically investigate these influences, but none of these seems to have had a lasting impact. Today we are still not much further with respect to preference change than we were at the middle of the last century.

This anthology hopes to provide a new impulse for research into this important subject. In particular, we have chosen two routes to amplify this impulse. First, we stress the use of modelling techniques familiar from economics and decision theory. Instead of constructing complex, all-encompassing theories of preference change, the authors of this volume start with very simple, formal accounts of some possible and hopefully plausible mechanism of preference change. Eventually, these models may find their way into larger, empirically adequate theories, but at this stage, we think that the most important work lies in building structure. Secondly, we stress the importance of interdisciplinary exchange. Only by drawing together experts from different fields can the complex empirical and theoretical issues in the modelling of preference change be adequately investigated.

Based on these ideas, we organised a 2-day workshop 'Models of Preference Change' at the Freie Universität Berlin in September 2006. We invited philosophers, logicians, economists and psychologists, and were happy to find many interested members of the audience engaging in illuminating discussions. This workshop was kindly sponsored by the Deutsche Forschungsgesellschaft, the Gesellschaft für Analytische Philosophie and the Philosophy Division of the Royal Institute of Technology, Stockholm. We thank these institutions for their support.

After the workshop, we decided to publish an anthology. We chose some of the workshop contributions, and invited four new contributors. We thank the editor-in-chief of the *Theory and Decision Library*, Julian Nida-Rümelin, for his kind invitation, and Springer for handling our project well. Thanks are also due to Kirsi L. Reyes for her help in formatting the document. Special thanks go to the referees of the contributed papers: Richard Bradley, John Cantwell, Peter Dietsch, Eduardo Fermé, Artur d'Avila Garcez, Patrick Girard, Natalie Gold, Conrad Heilmann, Aki Lehtinen, Fenrong Liu, Ben McQuillin, Martin Peterson, Odinaldo

Rodrigues, Jan-Willem Romeijn, Giacomo Sillari, Oliver Roy, Hannu Vartiainen, and Alex Voorhoeve. Their questions and criticisms helped to improve the papers of this volume considerably.

Stockholm and Helsinki
November 2008

Till Grüne-Yanoff
Sven Ove Hansson

Contents

Contributors

George Ainslie is Chief of Psychiatry at the Coatesville Veterans Affairs Medical Center, and Clinical Professor of Psychiatry at Temple University. Ainslie originally discovered hyperbolic discounting as an aspect of a broader empirical principal, Herrnstein's matching law. He has applied it to topics in economics, behavioral psychology, and the philosophy of mind in *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person* (Cambridge, 1992) and *Breakdown of Will* (Cambridge, 2001). His has published in journals like *Science*, *Psychological Bulletin*, *American Economic Review*, *Behavioral and Brain Sciences* and many others.

Luc Bovens is Professor of Philosophy at the London School of Economics and Political Science. His interests include Ethical Theory, Philosophy of Economics, Philosophy of Public Policy and Rational Choice. He has published a book on *Bayesian Epistemology* (Oxford, 2003, with Stephan Hartmann), and numerous articles in journals like *Mind*, *British Journal of Philosophy of Science*, *Social Choice and Welfare*, *Philosophy of Science*, and many others.

Richard Bradley is Professor of Philosophy at the London School of Economics. He does research in decision theory, hypothetical reasoning and the logic of conditionals, and has a special interest in the foundations of both Social Choice and Game Theory. He has published articles on these topics in *Philosophy of Science*, *Social Choice and Welfare*, *Erkenntnis*, *Synthese* and *Theory and Decision*, amongst others. His long-term research goal is the understanding the structure and dynamics of different types of rational social interaction.

Dick de Jongh is a Dutch logician and mathematician. He received his Ph.D. degree in 1968 from the University of Wisconsin-Madison under supervision of Stephen Kleene with a dissertation entitled *Investigations on the Intuitionistic Propositional Calculus*. De Jongh is mostly known for his work on proof theory, provability logic and intuitionistic logic.

Till Grüne-Yanoff is a Fellow of the Collegium of Advanced Study at the University of Helsinki. His research focuses on the methodology of economic modelling, on decision and game theory, and on the notion of preference in the social sciences. He has published in journals like *Synthese*, *Erkenntnis*, *Theoria*, *Journal of Economic Methodology*, amongst others.

Werner Güth is director of the Strategic Interaction Group at the Max Planck Institute of Economics in Jena. Before that he was professor of economic theory at the University of Cologne, the University of Frankfurt (Main) and Humboldt-University of Berlin. His main research topics are game theory, experimental economics and microeconomics, on which he has written seven books and over 160 articles. He considers himself as a social scientist with strong interests in psychology, philosophy (evolutionary) biology and the political sciences.

Sven Ove Hansson is professor in philosophy and head of the Department of Philosophy and the History of Technology, Royal Institute of Technology, Stockholm. He is editor-in-chief of *Theoria*. His research areas include value theory, decision theory, epistemology, and belief dynamics. He is the author of well over 200 articles in refereed journals. His books include *A Textbook of Belief Dynamics*, *Theory Change and Database Updating* (Kluwer, 1999) and *The Structures of Values and Norms* (CUP, 2001).

Brian Hill is Affiliate Professor at HEC School of Management, Paris, and an associate member of the Institut d'Histoire et de Philosophie des Sciences (IHPST). His main fields of research are in philosophy and decision theory; recent interests include belief change, conditionals, awareness and state-dependent utilities.

Hartmut Kliemt is Professor of Philosophy and Economics at the Frankfurt School of Finance and Management. Before, he held a professorship at Duisburg University. His main fields of research are Political Philosophy, Philosophy of Economics, Medical Ethics and Health Economics.

Fenrong Liu is Associate Professor of Logic at the Department of Philosophy, Tsinghua University, Beijing, China. She wrote her Ph.D. on Preference Dynamics and Agent Diversity at the Institute for Logic, Language and Computation of the University of Amsterdam in 2008. Her main research interests include dynamic preference logic, belief revision, multi-agent system, and bounded agency. She has published in journals like *Synthese*, *Journal of Logic, Language and Information*, *Journal of Applied Non-Classical Logic* and others.

Edward F. McClennen is Professor of Political Philosophy, at Syracuse University, and sometime Centennial Professor at the London School of Economics and Political Science. He specializes in Decision and Game Theory, Philosophy of Political Economy, Social and Political Philosophy. His publications include *Rationality and Dynamic Choice: Foundational Explorations* (CUP, 1990), "Pragmatic Rationality and Rules", *Philosophy & Public Affairs*, 26 (1997), and "An Alternative Model of Rational Cooperation," in Fleurbaey, M., M. Salles, and J. Weymark, Eds., *Justice, Political Liberalism and Utilitarianism* (CUP, 2008). He has also just finished a monograph, to be entitled, *Rational Society: Foundational Explorations*.

Stefan Napel holds the Chair of Microeconomics at the University of Bayreuth, Germany. His research interests include game theory, especially bargaining, measurement of power, and evolution; political economy of the European Union;

inequality and social mobility; and industrial organization. He has published in journals like *Journal of Economic Theory*, *Games and Economic Behavior*, *Economic Journal*, *Theory and Decision*, and many others.

Wlodek Rabinowicz is Professor of Practical Philosophy at Lund University. He was adjunct professor at the Research School for Social Sciences (RSSS), Australian National University, 2002–2007, Leibniz Professor at Leipzig University, 2000, president of the European Society for Analytic Philosophy, ESAP, 1999–2002, Member of Institut International de Philosophie, the Royal Swedish Academy of Sciences and the Royal Swedish Academy of Letters. Editor of *Theoria* and a former editor of *Economics and Philosophy*. Author of *Universalizability. A study in morals and metaphysics* (Reidel, 1979) and of numerous articles in moral philosophy, decision theory and philosophical logic in journals as *Journal of Philosophy*, *Theory and Decision*, *Synthese*, *Erkenntnis*, *Philosophy of Science*, among others. His current areas of research are formal axiology and decision theory.

Wolfgang Spohn holds a chair in philosophy and philosophy of science at the University of Konstanz, after holding professorhips at the universities of Bielefeld and Regensburg. He has been editor-in-chief of *Erkenntnis* and is a member of the Deutsche Akademie der Naturforscher Leopoldina. His research interests are epistemology and philosophy of science, in particular induction and causation; metaphysics; philosophy of language and philosophy of mind; logic, philosophical logics, and philosophy of logic and mathematics; decision and game theory, and the theory of practical rationality in general. He published the book *Grundlagen der Entscheidungstheorie* (1978) and over 70 articles in these areas, some of which are collected in his recent *Causation, Coherence, and Concepts, A Collection of Essays* (Springer, 2008).

Johan van Benthem is University Professor of logic and its applications, University of Amsterdam, and Henry Waldgrave Stuart Professor of philosophy, Stanford University. In the 1990s, he was a founding director of the Institute for Logic, Language and Computation ILLC in Amsterdam, a joint venture of mathematics, computer science, philosophy, and linguistics, for studying the structure and flow of information. His current main interests are logical dynamics of information, and interfaces between logic and games. He is a recipient of the national Spinoza Award, a member of the Royal Dutch Academy of Arts and Sciences (KNAW), the Academia Europaea (AE), and the Institut International de Philosophie (IIP).

Chapter 5
Why the Received Models of Considering Preference Change Must Fail

Wolfgang Spohn

Abstract First, the paper discusses the extent to which preference change is a topic of normative rationality; it confirms as one main issue the economists' search for a rational decision rule in cases in which the agent himself envisages to have changing preferences. Then it introduces so-called global decision models and shows that all the received economic models for dealing with preference change have that shape. The final section states two examples for global decision models, one with extrinsic, belief-induced and one with intrinsic preference change, and interprets each of them in two different scenarios in which different strategies are intuitively reasonable – the point being that global decision models cannot provide sufficient information for stating adequate decision rules. What the missing information might be is at least indicated at the end.

In this brief paper I want to give a specific argument for the title thesis. It is an entirely negative one, as far as it goes, unless one says it is positive to know how not to do things. A really positive treatment of the issue is, as far as I see, a very demanding and involved and as yet untold story.[1]

The title thesis seems ill expressed; either "of" or "considering" should be deleted. This would be an error, though. In order to understand why we have to briefly and generally discuss in which way preference change could be a philosophical topic at all; this is the task of Section 5.1. Having thus identified our topic, i.e., models of considering preference change, Section 5.2 introduces local and global decision models, as I call them, and explains that the latter are the received way of dealing with considering preference change. Section 5.3, finally, puts forward my negative argument: global decision models do not contain all items or distinctions that are intuitively required for rational decisions facing preference change.

W. Spohn
Department of Philosophy, University of Konstanz, 78457 Konstanz, Germany
e-mail: wolfgang-spohn@uni-konstanz.de

[1] I am indebted to Till Grüne-Yanoff and two anonymous referees for suggesting various improvements and clarifications.

T. Grüne-Yanoff and S.O. Hansson (eds.), *Preference Change: Approaches from Philosophy, Economics and Psychology*, Theory and Decision Library A 42,

5.1 Why Preference Change is a Philosophical Topic

To begin with, preference change is an indubitable fact. It is a complex phenomenon with multifarious possible causes. I prefer means because of aims; thus, information can change my preferences because it shows me that my aims are better reached by other means. My desire for food, i.e., hunger, changes several times a day because of food and digestion. I am getting tired of things. I am caught up by other things. I am maturing and aging, and my complex of aims, motives, desires, preferences, utilities changes accordingly. Whoever has kids knows that getting them motivated or sometimes also de-motivated is about the most difficult and imperspicuous part of educational work. Motivational and developmental psychologists have to tell a lot about this still very incompletely understood phenomenon.

What has philosophy to do with all this? As an empirical matter of fact, preference change may be hoped to be taken care of well by the human sciences from neurobiology over psychology up to social and political sciences. This is presumably not the task of philosophy, although philosophers can certainly assist in conceptual issues that abound in this area.

Besides, philosophy has a special competence in normative issues broadly understood. Introducing the normative perspective besides the empirical one makes things quite complicated. Roughly, we humans are receptive for normativity. Hence, the normative also serves as an empirical ideal that is often approximated by empirical facts; and reversely the empirical facts may often be taken as a prima facie indicator of the normative ideal.[2] The neat separation of the two perspectives does not work.[3] For this reason, normative philosophizing cannot leave empirical issues simply to the empirical human sciences, just as philosophy must listen to those sciences in pursuing normative questions.

Let us, however, ignore these complications and simply consider the normative perspective by itself. What can it say about preference change? This is not so obvious. Perhaps we should first distinguish two aspects of normativity, the rationality and the morality aspect; we *should* be rational, we *should* be moral, and these seem to be two different issues. (I wonder, though, how exactly to draw this distinction within the realm of normativity; it may turn out spurious in the end.)

So, let us more specifically ask: What is rational about preference change? There is a clear partial positive answer. Beliefs and desires, cognitive and conative attitudes are tightly entangled. I have already mentioned the most primitive instance, the practical syllogism: We have a goal; we believe certain means to reach the goal; therefore we want to take the means. We may call the desire for the means an extrinsic desire; there is nothing attractive in the means as such. In fact, the entanglement can take much more complicated forms, as decision theory teaches.

[2] For instance, the observation that people tend to divide fairly in the ultimatum game suggests that this behavior is rational and normatively required and that normative theories telling otherwise are false.

[3] In Spohn (1993) I have attempted to sort out this entanglement of the normative and empirical perspective; Spohn (2007) is a much briefer and sharper attempt.

Still, the point I want to note is clear already from the simple case: One's extrinsic desires, motives, preferences depend on one's (more or less firm) beliefs; if these beliefs change, the extrinsic desires change; and to the extent the former is rational, the latter is rational, too.

This point is, I think, well taken care of in the literature (though certainly not exhausted). The paradigmatic representation of extrinsic desires is given by expected utilities; the expectation of utilities relative to subjective probabilities is the paradigmatic account of the belief-desire entanglement. Moreover, we have clear and well-argued accounts of rational belief change and in particular of the rational change of subjective probabilities. Of course, decision theorists were always aware of the interaction of the two accounts. So, I do not want to bother here about this aspect of rational preference change.

Let us therefore continue to ask: What is rational about intrinsic preference change (which by definition cannot be accounted for in the way just discussed)? Now we are entering entirely insecure terrain. Most would say that intrinsic preferences or utilities are somehow given and not to be assessed as rational or irrational; hence their change is neither to be so assessed. Kusser and Spohn (1992) is one of the few attempts to overcome this negative attitude and to provide an extended notion of practical rationality. This is a minority position. For, those rejecting the proverbial *de gustibus non est disputandum* and accepting normative dispute over intrinsic preferences mostly tend to say that this is a dispute not about rationality, but about morality. So, if our philosophy somehow allows us to classify intrinsic preferences as (more or less) good or virtuous or morally acceptable, we automatically have a normative grip on intrinsic preference change: Changes towards the approved attitudes are good and should be supported, whereas changes in the reverse direction are bad and should be prevented. This is the rich field of moral education.

Here, I do not want to take a stance towards these difficult matters. I admit I belong to the patronizing camp (though with the appropriate bad conscience), and I even believe that intrinsic preference change can be assessed as being rational and not only as being moral. But I shall not further dwell upon these most important issues (since they are insecure and would take a much more elaborate discussion).

So, nothing seems left to talk about? No, we have not yet exhausted the rationality side of preference change. So far, we have only considered actual preference changes that may or may not be normatively and in particular rationally assessable. However, we can and must also consider foreseen preferences changes, raising the issue what practical rationality amounts to when one envisages changing preferences. So, our task now is not to assess some person's preference change by ourselves – we have put this to one side – but rather to assess a person's behavior that tries to take account of her possibly changing preferences (which we do not assess and she may or may not assess).

This is a problem decision and game theorists have always been aware of. If the considered preference change is of the extrinsic kind due to receiving information, standard accounts of strategic decision making well take account of it. And starting with Strotz (1955/56) there is a slowly growing literature dealing also with considering intrinsic or, as economists preferred to say, endogenous preference change. Let me just mention the oldest prototype of this kind of problem: Ulysses predicting

unwanted endogenous preference change under the influence of the songs of the sirens and thus rationally taking precautionary measures against yielding to this influence. This example points to a host of difficult issues and at the same time to a host of literature remaining more or less tentative.[4]

Now my title thesis makes sense: I want to critically reflect not on models of actual preference change, but on models of how to rationally behave when facing possible preference changes. What I want to argue is that we even do not have the appropriate conceptual means for generally treating these kinds of problems. If this should be correct, it is no wonder that our dealings so far are unsatisfactory. I want to argue this by working up to an example, and in fact to a recipe for constructing examples, which present two decision situations that are formally equivalent according to all models proposed for such problems, but clearly differ in their intuitive conclusions. If such examples are successful, they show that something is missing in all these models, and even though I have announced not to reach more positive results, the examples will at least point to what kind of information is missing. This is the program for the rest of the paper.

5.2 Local and Global Decision Models

So, what is the received modeling of envisaged preference change? We certainly have to focus on the decision and game theoretic representation of decision situations, i.e., on representing cognitive attitudes by subjective probabilities and conative attitudes by subjective utilities. Lots of variations in these representations are circulating, each variant responding to problems of another variant. For each variant, the problem of preference change poses itself in a different non-trivial disguise. However, all these variations are in quite a tentative state.[5] Hence, no experiments in this respect! I suppose my observations generalize to all the variant representations.

This point being fixed, how can decision situations considering preference change be modeled? A first step is to define $\langle i, S_i, P_i, U_i \rangle$ to be a *local decision model*, as one might call it, that consists of an agent i at a certain time, the set S_i of the agent's options of which he has to take one at that time, the agent's probabilities P_i for the relevant states of the world, propositions, or whatever, and the agent's utilities U_i for the relevant possible consequences, propositions, or whatever the precise construction is. Then, some *local decision rule* will say which options from S_i are optimal relative to P_i and U_i, under the assumption that $\langle i, S_i, P_i, U_i \rangle$ is a complete representation of (the relevant aspects of) the agent's decision situation; and if the agent is rational he chooses an optimal option. Usually, the local decision

[4] Elster (1979, 1983) is full of beautiful examples and problems. McClennen (1990) still seems the most advanced theoretical effort to systematically cope with these kinds of problems; see also the many references therein.

[5] See, e.g., Halpern (2003, Chapter 5) for some variant formal formats for cognitive and conative attitudes.

rule will be to maximize expected utilities that can be derived for S_i from P_i and U_i. For our context, however, the specific local decision rule is not really important. The important point about local decision models is only that P_i and U_i somehow capture everything relevant for determining locally optimal options, i.e., that the local decision rule operates only on P_i and U_i.

Local decision models are but a first step; changing preferences cannot be represented in them. For this purpose we have to consider whole evolutions of local decision models, or rather possible evolutions or trees, i.e., structures that I shall call here *global decision models*. Such a structure consists of a set N of nodes arranged as a tree. N tripartites into a non-empty set I of agents or agent nodes, a possibly empty set C of chance nodes, and a non-empty set E of end nodes, where the origin of the tree is an agent node and where the agent and the chance nodes have at least two successors and the end nodes have none. Finally, a local decision model $\langle i, S_i, P_i, U_i \rangle$ is associated with each agent node $i \in I$, where the set of options S_i is the set of successors of i (i.e., each option leads to a successor), P_i gives a subjective probability distribution over the successors of each chance node in C, and U_i is a utility function over all end nodes in E.[6]

The idea here is that the agent in the origin of the tree makes a choice, then or perhaps thereby and perhaps through the mediation of one or several chance nodes the situation moves to one of the subsequent agents whose probabilities and utilities may differ in *arbitrary* ways even over their common domain, and so forth till an end point is reached. Thus, a global decision model looks like a standard decision tree, the small, but crucial difference being that the action nodes of a decision tree representing only the options available at that node are replaced by agent nodes and thus by full local decision models. And precisely because these local models may contain arbitrarily varying probabilities and utilities such a global model is able to represent foreseen or envisaged extrinsic and intrinsic preference change. In the next section I shall introduce specific examples.[7]

Global decision models correspond to games in agent normal form as first introduced by Selten (1975; cf., e.g., Myerson 1991, Section 2.6). This model has proved to be useful in several game theoretical contexts. In order to fully understand it, one has to be clear about what an agent is. In philosophical terms, an agent is a possible stage of a person, or a player in a certain decision situation, so that different decision situations ipso facto contain different agents (that may constitute the same person or player, but the latter simply do not figure in the agent normal form). The suggestion, which we shall contest below, is that it suffices to consider agents in that dynamical context: Each agent simply tries to make the best out of his situation (when it is his turn – which may well not be the case since all the agents except those on the actually evolving branch remain mere possibilities).

[6] Alternatively, one might restrict P_i to the sub-tree originating at i or extend it to the agent nodes in the past of i. Each such detail is significant in principle, but not in the present context where we may leave them open.

[7] I want to avoid overformalization and think that global decision models as just characterized will do for our present purposes. If one really attempts to get formally explicit, things get quite complicated and look, e.g., as described in Spohn (2003, Section 4.3).

What the best is in each case need not be determined by a local decision rule referring at each agent node only to the associated local model. It may well be determined by a *global decision rule* that may be much more sophisticated. For instance, the agents may choose a Nash equilibrium or some other or stricter kind of equilibrium, and we may back up such a rule, which indeed refers at each local agent node to the entire global model, by assuming common knowledge of rationality and of the global decision situation among the agents. Again, though, the precise form of the global decision rule does not really matter. The crucial issue rather is whether a global decision model contains everything for reasonable global decision rules to operate on.

The view that this is indeed so seems to be commonly agreed among economists. It is particularly explicit in the global decision rule of so-called *sophisticated choice* that dominated the discussion since Strotz (1955/56). The basic idea of this rule is simple: The final agents of a global model (i.e., the agents with no further agent nodes between them and the endpoints) really face only a local decision situation; their situation is no longer multi-agent, strategic, reflexive, or whatever. So, a local rule will already tell what they will do. Assuming common knowledge of the global model, the predecessors of the final agents will therefore know what the final agents will do (if it will be their turn), and given this knowledge the predecessors can again locally optimize. Thus, backwards induction rolls back the global model from the endpoints to the origin.

This rough description hides many technical niceties. In order to overcome some of them, Peleg and Yaari (1973) introduced a game theoretic view on sophisticated choice and proposed the already mentioned global decision rule of a Nash equilibrium among the agents.

Strotz (1955/56) still did without chance nodes because he considered the simpler case of endogenous preference change foreseen with certainty (and because he was particularly interested in displaying the fatal consequences of myopia). However, one may also eliminate the chance nodes by assuming expectations with respect to the chance nodes to be implicitly contained in expected utilities. This is what Hammond (1976) does, the by then most general treatment of the issue; he assumes a global decision model without chance nodes and with arbitrary preference relations (instead of expected utility functions) attached to each agent node.

McClennen (1990), still the most careful treatment of the topic, also keeps his entire discussion within the confines of global decision models or equivalent formulations. Even in more recent surveys such as Shefrin (1996) and von Auer (1998, Part I) I do not find any tendency to transcend the frame of global decision models. These references may be sufficient evidence for my impression that it is indeed a common assumption that global decision models contain all information required for stating adequate global decision rules; the received models dealing with preference change have the shape of global decision models or an equivalent shape.

What is wrong with this assumption? One hint is provided by McClennen (1990). There, in Chapters 9 and 11, McClennen argues, convincingly in my view, that there is not only sophisticated choice, but also another reasonable global decision that he calls *resolute choice* (something mentioned, but not elaborated already by

Hammond (1976, pp. 162f.) under the label "precommitment"). Roughly, in resolute choice, the initial agent does not only take a choice in her decision situation, but fixes also the decisions of some or all the later agents; so, she does not let them decide from their point of view, but pre-decides or commits them to take a course of actions that is optimal from her point of view.

This description gives rise, by the way, to the observation that resolute choice does not make sense if the multi-agent setting is taken seriously, i.e., if the agents are independently deciding agents as they are assumed to be in a game-theoretic context. In that game-theoretic context, one agent cannot commit other agents. In more technical terms, resolute choice violates separability (cf. McClennen 1990, Section 9.7). Thus, resolute choice presupposes that all agents, or at least the initial agent and all agents pre-decided or committed by her constitute one person.

This is in fact the only interpretation to make sense in our context of preference change. It is *one* person pondering how to act when facing changing preferences; preferences varying across persons are not our problem. Let us thus explicitly assume that all agents in a global decision model are possible stages of *one* person. However, this assumption by itself does not change or enrich the conceptual resources of global decision models.

So far, resolute choice seems to be just another global decision rule so that one has to start an argument which of the global decision rules (mentioned or not mentioned so far) is the more or most reasonable. However, the problem presented by resolute choice is not just that it is a rival global rule forcing us into an argument over global rules. In my understanding, both, sophisticated and resolute choice, are reasonable global rules, depending on the case at hand; and the problem for global models is that they provide no means whatsoever for describing this dependence. Which parameters determine whether sophisticated or resolute choice or some other global rule is appropriate is not clear. The point is that global models as such, i.e. trees of local decision models (and chance nodes), do not contain these parameters. This will be clear from the examples to which I am about to proceed.

So, to be clear, these examples are intended as a criticism of the present state of the discussion about changing preferences that always proceeds, as far as I can see, within the confines of global decision models or essentially equivalent models. My claim is a bit vague since I refrained from developing the formal details. I am on the safe side, though, when I claim that my criticism will widely apply.

5.3 The Critical Examples

My examples will present two decision situations that are represented by the same global decision model, but intuitively require two different solutions. The examples thus suggest that global decision models are insufficient representations. I shall give two examples, one with an extrinsic, i.e., belief-induced preference change and one with an intrinsic preference change.

The first example is about agent 1 choosing from $S_1 = \{h_1, h_2\}$ and expecting a good or a bad outcome depending on the chance move with branches b_1 and b_2; let us more specifically assume.

$$U_1(h_1, b_1) = 2, U_1(h_1, b_2) = -2, U_1(h_2, b_1) = -10, U_1(h_2, b_2) = 2. \quad (5.1)$$

Thus, we are dealing with the following sub-tree T_1 (Fig. 5.1):

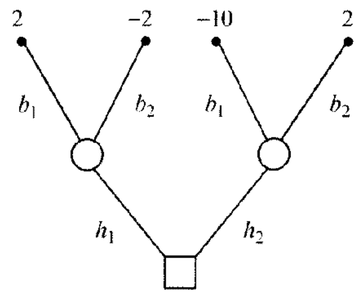

Fig. 5.1 Subtree T_1

The local model is still incomplete; it all depends on the probabilities. Let us assume $P_1(b_1) = P_1(b_2) = 0.5$ independently of the actions h_1 and h_2. Hence, $EU_1(h_1) = 0 > -4 = EU_1(h_2)$, and h_1 is the locally optimal choice.

The global model I want to consider now allows for an opportunity of belief change. So, agent 0 in the origin of the global model has the same utilities as agent 1, i.e., $U_0 = U_1$, and the same probabilities as far as the chance nodes in T_1 are concerned, i.e., $P_0 \supseteq P_1$. However, $S_0 = \{g_1, g_2\}$; that is, agent 0 has the option g_1 of refusing belief change, in which case he immediately turns into agent 1, i.e., moves to the sub-tree T_1, and he has option g_2 of allowing belief change that may take three different forms depending on the chance node C with three branches a_2, a_3, and a_4. Hence, the global model has the following form T_0 (Fig. 5.2):

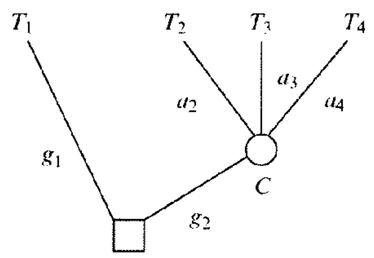

Fig. 5.2 Global model T_0

The global model contains five agents 0, 1, 2, 3, 4, each agent k being characterized by the (sub-)tree T_k. All of the agents 1, 2, 3, and 4 face the same decision; hence, $T_1 = T_2 = T_3 = T_4$ and $U_1 = U_2 = U_3 = U_4$. Only their probabilities may differ. Let us assume that agent 2 becomes certain of b_1, agent 3 becomes certain of b_2, and agent 4 still has equal probabilities for b_1 and b_2:

$$P_2(b_1) = 1, P_3(b_1) = 0, P_4(b_1) = 0,5,$$
$$P_2(b_2) = 0, P_3(b_2) = 1, P_4(b_2) = 0,5. \quad (5.2)$$

Hence, h_1 is optimal for agents 2 and 4 (as for agent 1), whereas h_2 is optimal for agent 3. The only information missing is the probabilities of agent 0. Suppose

$$P_0(a_2, b_1) = P_0(a_3, b_2) = P_0(a_4, b_1) = P_0(a_4, b_2) = 0.25, \text{and}$$
$$P_0(a_2, b_2) = P_0(a_3, b_1) = 0, \quad (5.3)$$

so that indeed

$$P_0(a_2) = P_0(a_3) = 0,25, P_0(a_4) = 0,5, P_0(b_1) = P_0(b_2) = 0,5,$$

and

$$P_0(.|a_k) = P_k \text{ for } k = 2, 3, 4. \quad (5.4)$$

This completes the specification of the global model; since the expected utilities of agents 1, 2, 3, and 4 differ, it is a model envisaging (extrinsic) preference change. Are we now in a position to tell what agent 0 should rationally do? No. I have two very different stories substantiating the formal figures.

In the first story, I have (b_1) or do not have (b_2) a serious disease requiring a special treatment (h_1) that works well and is harmless for those having the disease, but has quite unpleasant side effects for those not having it. This should make the utilities $U_0 = \ldots = U_4$ plausible. According to a preliminary check-up there is a good chance that I have that disease; thus, say, $P_0(b_1) = P_0(b_2) = 0.5$. The doctor informs me that there is a test the costs of which are negligible and that might tell more; there is a 50% chance of reaching certainty about the disease, with equal chances for positive (a_2) and for negative (a_3) certainty, and a 50% chance that the test remains mute (a_4). It is obvious how to judge this case: it would be silly to refuse the test (g_1) and to unconditionally decide for the treatment (h_1); rather I should undergo the test (g_2) because there is some chance of moving to T_3 and avoiding an unnecessary and unpleasant treatment (h_2).

Here is the second story. I have to catch a train at the other day that, as far as I know, might leave early, 8 a.m. (b_1), or late, 11 a.m. (b_2). So, I might go early to the station (h_1) running the risk of waiting for 3 h, or I might go late (h_2) and possibly miss the train. Again the distribution of utilities $U_0 = \ldots = U_4$ over the pairs $(h_i, b_j)(i, j = 1, 2)$ seems plausible. Now, for some reason I cannot get more information about the train; I am stuck with my uncertainty $P_0(b_1) = P_0(b_2) = 0.5$. In fact, it is even worse. I may, almost effortlessly, write up the two possible departure times (g_1), thus recalling them the next morning. Or I may not do so (g_2). In that case I know – I am not so young any more – that at the other morning I may well have forgotten that there are two possible departure times. Suppose there is a 50% chance of not forgetting (a_4), and a 50% chance of forgetting one departure time and thus becoming convinced of the other (a_2 or a_3) (where each of the two times has an equal chance to be forgotten). This is certainly not too artificial a scenario, and it is represented precisely by the global decision model specified above. However, I take it to be obvious that it is rational for agent 0 (me) to write up the two possible departure times (g_1), to thus preserve the uncertainty over night and

to leave early (h_1) instead of running the risk of getting opinionated the wrong way (through forgetting about the alternative) and missing the train.

Hence, we have here one global decision model considering extrinsic preferences, i.e., expected utility change and two different scenarios represented by the same global model, but with diverging intuitive rationality assessments. If this example is acceptable, there can be no adequate global decision rule operating on global decision models as explained.

Note that the first story about the disease involved learning (via the additional test), that probabilistic learning works by conditionalization, and that therefore, with respect to h_1 and h_2, P_0 had to be the mixture of P_2, P_3, and P_4 weighted by the probabilities of getting, respectively, into P_2, P_3, and P_4; my present probabilities always are the expectations of my better informed future probabilities. This is the so-called principle of iterability equivalent to van Fraassen's reflection principle – cf. Hild (1998). Therefore, I had to construct the second story in a way conforming to this principle as well, by accident, as it were. Given this construction, simply looking at the changing probabilities the process of possible forgetting could just as well have been a process of learning by conditionalization; this was the gist of the example. Of course, forgetting usually does not behave in this way. But it does in my story, and in not too forced a way, I think. Thus it serves my aim.

My second example considering intrinsic preference change is much simpler (and inspired by my recent travel experiences). Agent 0, i.e., I presently, has two choices, h_1 and h_2, and prefers h_1 over h_2; say, $U_0(h_1) = 1$ and $U_0(h_2) = 0$, though the numbers do not really matter. The choice need not be immediately made; so, agent 0 has two options, a_1 and a_2. He may either preserve his preference (a_1), thus turn into agent 1 with $U_1 = U_0$, and then choose h_1. Or he may try or test his preference (a_2), thus leaving it to (equal) chance (according to P_0) whether as agent 2 he preserves his preference ($U_2 = U_0$) or whether as agent 3 he changes it so that $U_3(h_1) = 0$ and $U_3(h_2) = 1$. Thus, we have the following global decision model (Fig. 5.3):

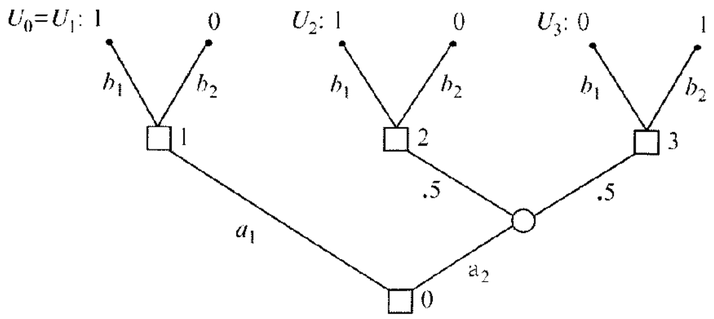

Fig. 5.3 The second global model

It is obvious what agents 1, 2, and 3 should do. But what should agent 0 do? Again, we have two different stories underlying the model.

In the first story, I am presently studying a beautifully made brochure by a first-rate travel agency, and I am immediately taken to a certain proposal; it looks

gorgeous and absolutely worth its price of €3,000. However, I cannot immediately order it (say, it's late in the evening). So, I may either commit myself (a_1) to immediately going to the agency the next morning (say, simply by building up determination and not allowing further doubts). Or I may sleep on the matter for a night (a_2) and see whether my present excitement keeps on, being unsure whether it really does. What is the reasonable thing to do in this case? I do not think that there is any objective answer. However, one reasonable attitude I might take (and which many will share) is that I mistrust the seductive power of such brochures, mistrust my seducibility, and thus choose to sleep on the matter (a_2).

In the second story, I walk through a picturesque street of a foreign city in which street hawkers offer the cheap, but ornate goods typical of their country. Initially, I think the goods are never worth the €20 for which they are offered and not even the €5 at which the bargain might end; so initially I prefer not buying (b_1) to buying (b_2). However, the dealers can be quite obtrusive, and I have to develop a strategy before walking down the street. Either, I close my mind (a_1), determinately not paying attention to the dealers (who are not the sirens, after all), and thus stick to my initial preference; or I have an ear for them (a_2), risking that they talk me into reversing my preference and buying their stuff. Again, I do not think that there is an objectively recommended attitude. This time, though, one may plausibly be determined not to buy any of the junk and conclude that it is reasonable to ignore the dealers (a_1).

The point of the example is the same as before. There is a global model considering preference change, indeed an intrinsic one, since it is directly the attraction things exert on me that changes and not any information I have about them. Yet, there are two different scenarios substantiating this model, and one would like to be able to rationalize different courses of actions for these scenarios. However, the global model cannot provide the means for doing so.

The construction recipe of these examples is obvious; so one can think of many variations. One may argue about the adequacy of the formal representations of such examples. Such arguments are painfully undecidable, though, and one may therefore distaste debates on this intuitive level. It is, however, impossible to avoid such debates. Normative theory by itself cannot decide what is rational; it lives from being in reflective equilibrium with our intuitions about what is reasonable and what is not.

One may seek for more fine-grained formal representations of the examples that keep within global decision models, but show a difference in each critical pair. I admit that this might be done even with the above examples in a plausible way. One may counter, though, with more sophisticated examples in which the old problems return. And so on. The ensuing race of sophisticated formalizations and counterexamples is again hardly decidable. I would like to block such considerations by an invariance principle, as I have called it, which I have stated and defended in an entirely different context, but which applies in this context as well; cf. Spohn (2009, Chapter 16).

I rather conclude from my examples that global decision models are indeed incomplete. No generally acceptable global decision rule can be stated on that level.

I also find that the examples clearly suggest what is missing in the global models. The crucial parameter missing is, it seems to me, whether the evolution of local decision situations leads to what one might call superior or inferior local situations. Superiority and inferiority need not be objectively fixed. Each person, however, has a judgment about this when surveying the evolution of local situations. When she learns something, she can make a better informed decision. When she forgets something or is not at her cognitive height for some other reason, she is in a worse position for deciding. So she is when she is in an emotional turmoil or about to be seduced or more seriously irresponsible, whereas a sober state is apt for better decisions. Or she may reversely have learnt to listen to her rare excitements and take its preservation to be subjectively superior to boring soberness. And so forth.

In any case, I believe that this was the crucial parameter governing the examples I have given and missing in global decision models. Proposing this conclusion is one thing. Constructively specifying how global decision models may be enriched by such a parameter and how global decision rules may be made to depend on it is, however, quite another and obviously much more complicated thing.

References

Elster, Jon. 1979. *Ulysses and the Sirens. Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

Elster, Jon. 1983. Sour Grapes. *Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Halpern, Joseph Y. 2003. *Reasoning about Uncertainty*. Cambridge, MA: MIT Press.

Hammond, Peter J. 1976. Changing Tastes and Coherent Dynamic Choice. *Review of Economic Studies* 43: 159–173.

Hild, Matthias. 1998. Auto-Epistemology and Updating. *Philosophical Studies* 92: 321–361.

Kusser, Anna and Wolfgang Spohn. 1992. The Utility of Pleasure is a Pain for Decision Theory. *Journal of Philosophy* 89: 10–29.

McClennen, Edward F. 1990. *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.

Myerson, Roger B. 1991. *Game Theory. Analysis of Conflict*. Cambridge, MA: Harvard University Press.

Peleg, Bezalel and Menahem E. Yaari. 1973. On the Existence of a Consistent Course of Action When Tastes are Changing. *Review of Economic Studies* 40: 391–401.

Selten, Reinhard.1975. Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games. *International Journal of Game Theory* 4: 25–55.

Shefrin, Hersh M. 1996. Changing Utility Functions. In *Handbook of Utility Theory*, eds. S. Barbera, P. J. Hammond, and C. Seidl, 569–626. Dordrecht, The Netherlands: Kluwer.

Spohn, Wolfgang. 1993. Wie kann die Theorie der Rationalität normativ und empirisch zugleich sein? In *Ethik und Empirie. Zum Zusammenspiel von begrifflicher Analyse und erfahrungswissenschaftlicher Forschung in der Ethik*, eds. L. Eckensberger and U. Gähde, 151–196. Frankfurt a.M., Germany: Suhrkamp.

Spohn, Wolfgang. 2003. Dependency Equilibria and the Causal Structure of Decision and Game Situations. *Homo Oeconomicus* 20: 195–255.

Spohn, Wolfgang. 2007. The Core of Free Will. In *Thinking About Causes. From Greek Philosophy to Modern Physics*, eds. P. K. Machamer and G. Wolters, 297–309. Pittsburgh, PA: Pittsburgh University Press.

Spohn, Wolfgang. 2009. *Causation, Coherence, and Concepts. A Collection of Essays*. Dordrecht, The Netherlands: Springer.

Strotz, Robert H. 1955/56. Myopia and Inconsistency in Dynamic Utility Maximization. *Review of Economic Studies* 23: 165–180.

von Auer, Ludwig. 1998. *Dynamic Preferences, Choice Mechanisms, and Welfare*. Berlin: Springer.