Till Grüne-Yanoff
Sven Ove Hansson
Editors

Theory and Decision Library A

Preference Change

Approaches from philosophy, economics and psychology

Springer

PREFERENCE CHANGE

APPROACHES FROM PHILOSOPHY, ECONOMICS AND PSYCHOLOGY

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

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

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 *Till Grüne-Yanoff*
November 2008 *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 1
Preference Change: An Introduction

Till Grüne-Yanoff and Sven Ove Hansson

Abstract In this introduction, we discuss a number of reasons why preference change has been neglected in the social sciences, in particular in economics. We argue that recent developments make this neglect less acceptable than it may have been in the past. We then propose a modelling approach to preference change that starts out with the standard preference notion and pays careful attention to its formal properties, in particular the connections between preference relata, the logical constraints on preferences, and their temporal specification. Based on this proposal, we categorise preference change models into four groups: those that derive changed preference from more basic structures, those that refer to the temporal dimension, those that focus on consistency preservation, and finally those that offer an evolutionary account. Using this categorization, we also introduce the other papers of this anthology.

1.1 Why Investigate Preference Change?

1.1.1 Reasons for Neglect

In the formal social sciences, preference change has generally been given scant attention. This is particularly true for economics. At least three reasons for this neglect can be identified. First, there is a long tradition of 'division of labour' between economics and the other social sciences, and changing preferences have mainly been located on the non-economic side. Classical economists like John Stuart Mill conceived of political economy as investigating only one pervasive aspect of human action, namely that connected to the production of wealth (Mill 1844, p. 318).

T. Grüne-Yanoff (✉)
Helsinki Collegium of Advanced Studies
e-mail: till.grune@helsinki.fi

T. Grüne-Yanoff and S.O. Hansson
Royal Institute of Technology, Stockholm
e-mail: soh@kth.se

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

Changes in tastes resulting from education, etc., would not fall under this aspect. Hence, other social sciences would have to contribute to an explanatory synthesis of human behaviour, if preference changes were involved.

When the prevailing view changed to the Robbinsian definition of economics as a 'science of choice', this division of labour became even more pronounced. In the 1930s, the economist Lionel Robbins (1932) and the sociologist Talcott Parsons (1934, 1937, 1970) established a new consensus about the division of labour between their two disciplines. Economics was to focus on the rational choice of means to serve given ends, and sociology on the explanation of the social origins of those purposes or ends. This consensus survived with little challenge until the 1970s. Even today a significant number of social scientists would define the two subjects in these terms.

Interestingly, this division of labour became a defining feature of the disciplinary division itself. Neither economics nor sociology defined themselves in terms of distinctive and mutually exclusive sets of objects of analysis. Instead, they were separated in terms of core concepts and approaches to analysis. Economists would emphasize individual rationality. The framework of rational choice under constraint with given preferences was the defining feature of the discipline. Sociologists would emphasize the roles of structures, culture and – particularly relevant for the current discussion, values (cf. Hodgson 2008).

A second reason for the neglect of preference change was a conviction of many micro-economists that human preferences ultimately do not change. On a superficial level, people's desires may seem to vary, Stigler and Becker argued in their influential paper *De gustibus non est disputandum*. Yet upon closer inspection, tastes, the foundations of these desires, remain stable:

> [O]ne does not argue about tastes for the same reason that one does not argue about the Rocky Mountains—both are there, and will be there next year, too, and are the same to all men. (Stigler and Becker 1977, p. 76)

This position may be interpreted either as the ontological claim that preferences indeed are stable, or alternatively as the methodological claim that explanations based on stable preferences are better than those that refer to preference changes. The second interpretation can be based on the assumed relation between explanatory power and simplicity: explaining any conceivable human behaviour through the paradigm of individuals maximizing utility constrained by income and present capital stocks is simpler than supposing that tastes change.

The stability of tastes over time implied by the Stigler–Becker analysis was empirically supported in a number of studies. Landsburg (1981) studied meat consumption behaviour in England for the period 1900–1955 in an attempt to find counterevidence against the Strong Axiom of Revealed Preference. This axiom requires that whenever a bundle A_1 is chosen over a bundle A_2, and in another situation A_2 is chosen over a bundle A_3, and so on, then in a situation where A_1 and A_n are available, A_1 is always chosen over A_n. For the entire period Landsburg found no instances of such rejections. Similar results are found in a nonparametric study by Chalfant and Alston (1988) on Australian meat demand from 1962 to 1984 (see however Grüne-Yanoff 2004 for a critical discussion of the methods used in these studies).

A third reason for the neglect of preference change lies in the conviction of many macroeconomists that institutional change, in comparison to individual value change, is by far the more important explanatory factor of economic growth. Modern macroeconomists (as well as institutionalists like Douglass North) here echo both Adam Smith and Karl Marx: institutions made the difference, whether limited government, competition for profits, the expansion of markets, secure property rights, the enclosure of common lands, or empire. These changed institutions provided people with new incentives, and thus changed their behaviour. People's change in preferences or values, in contrast, need not be invoked in such explanations.

1.1.2 Rising Interdisciplinary Exchange

In recent years, two developments in economics and its neighbouring disciplines have contributed to a breakdown of the Robbins–Parsons division of labour. First, economics has expanded into subject fields beyond commodity consumption and monetary markets. Paradoxically, by advancing Robbins' non-subject-bound definition of economics, economists who ventured into these areas saw more of a need to engage with the formation of preferences. The work of economist Gary Becker is a paradigmatic example of this approach. When investigating family behaviour, the relation between crime and punishment, or discrimination in labour and goods markets, he left behind the narrow confines of assumed self-interest and instead based his explanations on a 'much richer set of values and preferences' (Becker 1993, p. 385). This explanatory project led to an increased focus on the variety of preferences and values, and the need to account for them theoretically. In Becker (1996) he offered such a theoretical account, arguing that past experiences and social influences form preferences and values. He applied these concepts to assessing the effects of advertising, the power of peer pressure, the nature of addiction, and the function of habits.

Secondly, economics not only expanded into neighbouring fields, it also increasingly imported concepts and ideas from other disciplines, especially psychology. This brought with it a wealth of evidence about preference instability. Social psychologists, for example, have found that human attitudes (including likes and dislikes, hence related to preferences) may be much less enduring and stable than has traditionally been assumed (Schwarz and Strack 1991; Tourangeau 1992). In cognitive psychology, numerous experiments have provided evidence of taste changes, especially in relation to perceived risk levels (Kahneman et al. 1982), and in response to changing constraints and abilities (Aronson 1972).

Not only did psychology provide evidence of preference instability, it also offered theories why preferences change. Social psychologists, for example, have long argued that social influence is an important determinant of individual preferences (Deutsch and Gerard 1955; Nisbett and Ross 1980; Cialdini and Goldstein 2004). Cognitive psychologists have offered various non-standard decision theories involving context-dependent utilities. Most well-known amongst these

is Kahneman's and Tversky's prospect theory (Kahneman et al. 1982). In particular the research on cognitive biases has found its way into economics itself. Behavioural economists have investigated various cognitively 'anomalous' effects on preferences (Kahneman et al. 1991) and they have also investigated the affectual bases of human preferences (Loewenstein 1996, 2000; Loewenstein and Schkade 1999; Loewenstein and Angner 2003).

Outside of psychology, marketing and consumer researchers have offered theories about the genesis of tastes and preferences (Holbrook and Schindler 1989, 1994, 1996; Schindler and Holbrook 1993). Anthropologists also provide a wealth of evidence for preference change. Barry et al. (1959), for example, argue that there is a connection between forms of livelihood and patterns of child-rearing, with consequences on those children's preferences. Dreeben (1968) suggests that universal schooling has effects on individual values and preferences. Edgerton (1971) proposes a relation of livelihood and preference for independence. This anthropological literature has recently begun to attract attention from economists (Bowles 1998; see also Henrich et al. 2005 for collaboration between anthropologists and economists on preference variation and change). Thus, with the breakdown of the Robbins–Parsons divide in these two ways, the need for more rigorous models of preference change has increased.

1.1.3 Preference Endogeneity

Sociological and philosophical critics of economics have often invoked a relationship between economic structures on the one hand and values and tastes on the other. Works like *Capitalism, Socialism and Democracy* (Schumpeter 1942), *The Great Transformation* (Polanyi 1944) and *People of Plenty* (Potter 1954) argue that the growth of wealth and economic institutions have influenced (often in negative ways) the judgment and the values of people living and working under these conditions. Yet none of these authors have offered more precise, causal accounts of these influences.

Beginning in the 1950s, some economists tried to incorporate these effects in their demand–supply models. Two approaches can be distinguished. First, 'endogenous change in preferences' (Hammond 1976) or 'habit formation' refers to a situation in which what one consumes in the present alters the preferences one has in the future. A perspicuous example of endogenous preference formation is 'sodium hunger' (Schulkin 1991) – increased consumption of salty foods leads to increased taste for salty foods. Work on habit formation has mostly focused on demand systems with parameters that depend on the consumption history of individuals. Important early contributions are von Weizsäcker (1971) and Pollak (1976b, 1978).

Second, 'preference interdependence' refers to a situation in which what *others* consume in the present alters the preferences one has in the future. Preference interdependence was described by Adam Smith (1776, Bk I, Chapter XI) and Thorstein Veblen (1899). It became widely known as the 'bandwagon effect'

(Leibenstein 1950). Duesenberry (1949) offered evidence based on aggregate data to indicate the importance of preference interdependence. Further studies include Fisher and Shell (1972), Krelle (1973), Pollak (1977). Gaertner (1974), Pollak (1976a) and Hansson (2004), who investigate preference interdependence by letting the parameters of an individual utility function depend on the consumption of other individuals. Kapteyn and Wansbeek (1982) synthesize both approaches in their theory of preference formation, assuming that an individual's welfare function is dependent on the distribution of consumption patterns the individual has observed over time. This includes both the individual's own consumption and the consumption by others in his or her social reference group.

After the 1970s this research was largely abandoned by economists. A possible reason for this abandonment may have been the lack of cognitive models that would have allowed a better understanding of how preferences are affected by the behaviour of oneself and others. With the considerable advances in the cognitive sciences, the development of new models of preference change seems to be a worthwhile extension of these earlier theoretical projects.

1.1.4 Evolutionary Explanations of Growth

In recent years, the exclusivity of technology and institutional development as explanatory factors of growth has been questioned. As part of the influential 'unified growth' approach, which tries to offer a single theory explaining the transition from Malthusian stagnation to self-sustaining growth, it has been argued that changes in people's preferences and selective pressure on those preferences also contribute to growth. This idea goes back to the eighteenth century philosopher David Hume (Grüne-Yanoff and McClennen 2008). In an influential paper, Galor and Moav (2002) develop a full-fledged evolutionary growth theory on these premises. They argue that an upward drift in the quality of human populations was critical for the transition from 'Malthus to Solow'. In particular, it was not institutions but people that changed, and their new values – 'thrift, prudence, negotiation, and hard work' – led them to save, work, and invest in ways that would eventually bring about the industrial revolution (see also Clark 2007 for an expanded version of this argument).

This new approach to macroeconomic growth clearly presupposes that preferences change in specific ways. Modelling preference change is thus a prerequisite for a precise formulation of this explanatory account.

1.1.5 New Questions on Rationality

Contemplating the possibility of endogenous change as discussed in Section 1.1.3 inevitably leads to the question: what is the meaning of 'rational behaviour' in a setting where the act of consumption may induce a change in the consumer's

preferences vis-à-vis subsequent consumption? If individuals anticipate that their behaviour will affect their future preferences, this effect should be taken into account when rationally choosing between different options.

The sophisticated behaviour approach of Strotz (1955–1956) assumes that individuals know that their present choices influence their future preferences, and make rational choices based on this knowledge. In particular, sophisticated choosers anticipate which of their currently available options will lead to preference changes disadvantageous to them, and avoid choosing these options. This gives rise to a variety of problems of consistency, existence and stability of plans and choices over time (Pollak 1968; von Weizsäcker 1971; Peleg and Yaari 1973; Hammond 1976; Winston 1980; Laibson 1997; Edvardsson et al. 2009). In addition to these problems, this approach also presupposes that the decision maker knows sufficiently well how a current choice would affect future preferences. Alternatively, McClennen (1990) suggested that individuals will form intertemporal plans, and try to stick to their plans (with the help of external devices and/or internalised practices) even when preference reversal threatens at some later stage.

These accounts of intertemporal choice presuppose that the decision-maker is able to predict and influence her own future preference changes. They therefore make it necessary for decision theorists to develop models of decision makers' preference changes.

1.1.6 Questions About Welfare Measurement

The most common welfare measures of traditional normative economics are based on consumer preferences. A *Pareto improvement* consists in a change in goods allocation that leaves some individuals 'better off' with no individual being made 'worse off'. Here 'better off' is often interpreted as 'put in a preferred position'. An allocation is Pareto efficient if no Pareto improvement is possible. It is commonly accepted that outcomes that are not Pareto efficient are to be avoided, and therefore Pareto efficiency is an important criterion for evaluating economic systems and public policies. A second, broader criterion is *Kaldor–Hicks efficiency*. Under Kaldor–Hicks, an outcome is considered more efficient if a Pareto efficient outcome can be reached by arranging some compensation from those that are made better off to those that are made worse off. Again, both Pareto efficiency and compensation are commonly interpreted as 'being put in a preferred position'.

The use of these welfare criteria becomes problematic if preferences are unstable. Yaari summarises this concern aptly: 'What measuring-stick can one use to evaluate the performance of an economic system, now that consumers' preferences can no longer be used (because they keep changing) to construct an unambiguous measure of performance?' (Yaari 1977, p. 158). Beyond the technically-minded question how it would be possible to obtain a consistent welfare measure, the possibility of changing preferences also led to a broader social critique of economic objectives. Critics like Galbraith (1958) and Marcuse (1964) asked what the merit of establishing a

system designed to fulfil consumers' wants would be, given that these wants are themselves the products of corporate manipulation, through advertising and other means (cf. also Koopmans 1957, p. 166).

The implications of interdependent preferences and habit formation on welfare economics were studied by Duesenberry (1949), Harsanyi (1954), von Weizsäcker (1971), Fisher and Shell (1972) and Pollak (1976b). Hansson (2004) proposed a two-tiered model in which a person's well-being may depend on the material resources of other persons. Pareto efficiency on the level of well-being need not coincide with Pareto efficiency on the level of material resources. Under certain conditions, Pareto efficiency on the level of well-being will require non-Paretian inequality-reduction on the level of material resources.

One possibility of making meaningful welfare comparisons based on variable preferences, as suggested by Weisbrod (1977), is to apply the Pareto criterion twice, based on the initial and the new preferences. However, these considerations have found little acceptance in mainstream welfare economics. Again, the lack of models of particular mechanisms has limited the success of those concerned with the implications of preference change.

1.2 The Formal Preference Notion as the Basis of Models of Preference Change

1.2.1 The Need for Structured Models

If one accepts the evidence of preference instability, and also accepts that certain theories have to include preference change in order to be adequate, then the question arises how preference change is best introduced for the purposes at hand. Two methodological problems arise immediately. First, it is possible to explain almost anything on the unrestricted hypothesis that consumers' preferences are changing over time. The empirical power of discrimination of an economic theory based on the hypothesis of changing preferences is likely to be low, unless this hypothesis is furnished with sufficient structure. Second, with changing preferences, it may no longer be possible to explicate the term preference in terms of the consumer's potential acts of choice, and it may become necessary to rely instead on an attitudinal or introspective explication. Both attitudinal and introspective approaches are viewed with scepticism in the economics community. This may partly be due to intricate questions concerning their validity. However, current economics also prefers a theorizing style very different from that of inductive generalizations based on a set of observations. Thus, even if economists were more favourably inclined towards introspective evidence, the question would remain where to apply this evidence in economic models. Instead of more empirical evidence, thus, the first thing that economists need in order to incorporate preference change in their theories is an appropriate theoretical structure.

The papers in this anthology address these methodological concerns by developing various *models* of preference change. Modelling here means the development of

formalized possible mechanisms, either for the purpose of isolating certain features of the world, or of creating simplified hypothetical worlds whose investigation may lead to useful information about the real world. (On modelling methodology, see the papers in Grüne-Yanoff 2009.) These models are not meant to reflect the complexity of preference changes found in the real world. Rather, they concentrate on certain possible aspects of preference change, and develop the structure and dynamics of these aspects in ways that are hoped to elucidate possible causal and mechanistic structures of preference change.

1.2.2 The Standard Notion of Preference

The basis for all these modelling attempts is what we will call the standard notion of preference. Preferences are almost always assumed to have structural properties of a type that is best described with formal tools such as those used in preference logic, expected utility theory and set theory. Structural properties thus described are a suitable starting point for the development and categorization of models of preference change. In this section, we review the basic structural properties of the notion of preference, and point out their connection to different types of preference change.

A preference expresses a relational value judgment. It is relational in the sense that it connects two or more *relata*. These relata may be propositions expressing states of affairs, events, etc. or they may be bundles of goods. Preference is a value judgement in the sense that it compares relata with respect to (some aspect of) their value. There are two fundamental comparative value concepts, namely "better" (strict preference) and "equal in value to" (indifference). The relations of preference and indifference between alternatives are usually denoted by the symbols $\succ$ and $\sim$ or alternatively by P and I.

The relation "better than or equal in value to" (weak preference) is usually denoted by the symbol $\succeq$ or by R. It can be introduced disjunctively, so that $A \succeq B$ holds if and only if either $A \succ B$ or $A \sim B$ holds. In accordance with a long-standing tradition, $A \succ B$ is taken to represent "B is worse than A", as well as "A is better than B".

Particularly in economics it is common to base a preference model on a utility function u. This can be done with the two defining equations: (1) $A \succ B$ if and only if $u(A) > u(B)$ and (2) $A \sim B$ if and only if $u(A) = u(B)$. In the most common usage in the social sciences, preference judgments represent subjective judgments. However, an alternative interpretation in terms of objective betterness is compatible with the structure.

1.2.3 Relata

The objects of preference are represented by the relata of the preference relation (A and B in $A \succ B$). In order to make the formal structure determinate enough,

every preference relation is assumed to range over a specified set of relata. In most applications, the relata are assumed to be mutually exclusive, i.e. none of them is compatible with any of the others. Preferences over a set of mutually exclusive relata are referred to as *exclusionary* preferences (Hansson 2001a). The relata are often also called alternatives, and the set of relata is called the *alternative set*.

Preference change can be driven by changes in the alternative set. If relata are added or removed, then the preference relation will have to be changed accordingly. Furthermore, changes in the agent's beliefs about the relata can be drivers of preference change. New beliefs about an alternative can lead us to rank that alternative higher or lower than we did before. Such belief changes may or may not in their turn be caused by changes in the actual properties of the relata.

In philosophical treatments of preference logic, alternatives are commonly taken to be states, represented by sentences or propositions. In contrast, economics commonly conceives of alternatives as bundles of goods. They are represented by vectors, where each position in the vector represents a specific good, and the magnitude at that position denotes the quantity of that good. However, this representation involves a problematic ambiguity. For example, it is not coffee *per se* that one prefers to tea *per se*. Consumers may prefer *drinking* coffee to *drinking* tea, and merchants may prefer *stocking* coffee to *stocking* tea, etc. If preferences are subjective evaluations of the alternatives, then what matters are the results that can be obtained with the help of these goods, not the goods themselves.

Economists have tried to solve this ambiguity by coupling preferences over goods with household production functions (Lancaster 1966; Becker and Michael 1973). Philosophers have also contributed to this debate by distinguishing between different levels of preferences. On the most basic level, *exclusionary* preferences compare relata with maximal detail. From these, *combinative* preferences, which compare relata of lesser detail, are derived. In most variants of this approach, the underlying alternatives (to which the exclusionary preferences refer) have been possible worlds, represented by maximal consistent subsets of the language (Rescher 1967; von Wright 1972; Hansson 1996). The derivation of combinative preferences from exclusionary preferences can be achieved with a representation function. By this is meant a function f that takes us from a pair $\langle p, q\rangle$ of sentences to a set $f(\langle p, q\rangle)$ of pairs of alternatives (perhaps possible worlds). Then $p \succeq_f q$ holds if and only if $A \succeq B$ for all $\langle A, B\rangle \in f(\langle p, q\rangle)$ (Hansson 2001a, pp. 70–73). A change in the function f may then lead to a preference change. For example, I prefer being rich to being poor, because I prefer every way I may become wealthy to every lifestyle in which I stay poor. However, you then point out to me various lifestyles in which I remain poor, and which I prefer to the corresponding lifestyles in which I would become rich. Consequently, I abandon (or qualify) my preference for being rich (Grüne-Yanoff 2008).

Decision theorists have developed models in which the value (desirability) of a proposition is linked to the values (desirabilities) of the possible worlds in which that proposition is true. One common way to do this is to assign to each possible world a weight according to its probability. The desirability of a proposition p then

depends on the desirabilities and probabilities of the worlds w in which it is true, thus:

$$des(p) = \frac{\sum_{\{w \in W \mid p \in w\}} prob(w) \times des(w)}{\sum_{\{w \in W \mid p \in w\}} prob(w)}$$

where W is the set of possible worlds. They then argue that 'the desirability of a proposition is a weighted average of the desirabilities of the cases [worlds] in which it is true, where the weights are proportional to the probabilities of the cases' (Jeffrey 1983, p. 78). This is of course a generalized version of expected utility theory, well known to economists and decision theorists (Savage 1954). It provides us with an additional mechanism for preference change: a change in the probabilities may lead to a change in preferences.

1.2.4 Logical Constraints on Preference

In preference logic, preference axioms (postulates) are used as premises. Some of the most important of these axioms are:

1. $A \succ B \rightarrow \neg(B \succ A)$ (*asymmetry of preference*)
2. $A \sim B \rightarrow B \sim A$ (*symmetry of indifference*)
3. $A \sim A$ (*reflexivity of indifference*)
4. $A \succ B \rightarrow \neg(A \sim B)$ (*incompatibility of preference and indifference*)
5. $(A \succeq B) \wedge (B \succeq C) \rightarrow A \succeq C$ (*transitivity of weak preference*)
6. $(A \succ B) \wedge (B \succ C) \rightarrow A \succ C$ (*transitivity of strict preference*)
7. $A \sim B \wedge B \sim C \rightarrow A \sim C$ (*transitivity of indifference*)
8. $(A \succ B) \vee A \sim B \vee (B \succ A)$ (*completeness*)

(For more details on such properties, see Hansson 2001b; Hansson and Grüne-Yanoff 2006.) The status of some of these axioms is controversial. Even among scholars who hold a particular preference axiom to be plausible, opinions may differ about its status. There are least four options. First, a preference axiom can be *constitutive* of the notion of preference. This means that it is conceptually impossible for a person to hold preferences violating the axiom in question. Whatever it is that does not satisfy a constitutive axiom cannot be preferences. On the above list postulates (1–4) are obvious candidates for status as constitutive. Secondly, satisfaction of a preference axiom can be a necessary condition for preferences to be *consistent*. Thirdly, its satisfaction can be a necessary condition for *rationality*. In practice, the distinction between preference consistency and preference rationality is seldom made. Of the above axioms, in particular (5–7) have been treated as rationality requirements (but their status as rationality axioms has also been contested). Fourthly there may be pragmatic reasons for an agent to satisfy a particular axiom. Hence, it can be argued in favour of our axiom (8) that once you have developed a complete preference relation over a set of alternatives you are prepared to make any choice among the alternatives without having to reconsider the value issues at stake.

From the viewpoint of modelling preference change, if a preference axiom is considered to be constitutive, then preferences violating it should in principle not be representable in the preference modelling. If a preference axiom is considered to be a requirement of consistency or rationality, then preference states violating it should be treated in the same way as inconsistent belief states are treated in belief revision, namely as unsatisfactory intermediate states in need of immediate repair. Therefore, consistency (rationality) requirements can be drivers of preference change, which of course makes them particularly interesting in a preference change framework.

1.2.5 Temporal Specification

Preferences can be temporally specified in at least two different ways: Either the relata or the preferences themselves can be associated with specific moments in time.

In the formal framework, relata can be temporally specified in two ways. For concreteness, let us assume that the relata are states of affairs. On the first approach, temporal specification is part of the meaning of the basic representation of states of affairs. Hence, a relatum A can be taken to mean "Peter visits his mother at time t". On the second approach, the basic representation of states of affairs is timeless. In that case the temporal aspect has to be treated separately, most conveniently by forming pairs of such timeless states of affairs and points in time. Then a relatum B can be taken to mean "Peter visits his mother", and (B, t) means that this holds at time t. Clearly, A and (B, t) are synonymous expressions. The latter form has the important advantage of allowing for explicit treatment of temporal aspects.

The preference judgment itself can also be temporally specified. This can be expressed with a temporal index on the relation. The use of such an index does not decrease the need for temporal specification of the relata. It is quite possible that at t_3, C at t_1 is preferred to than C at t_2, whereas the contrary is true at time t_4. This can be expressed by the two statements $(C, t_1) \succ_{t_3} (C, t_2)$ and $(C, t_2) \succ_{t_4} (C, t_1)$. It is important in a precise discussion of the temporal aspects of preferences to distinguish between the temporal indexing of relata and of the preferences themselves. Hence, the statement "A is better than B at time t" can mean either $A \succ_t \mathrm{B}$, $(A, t) \succ (B, t)$, or $(A, t) \succ_t (B, t)$. Depending on which temporal indexing is used, a shift in the temporal dimension may or may not constitute a preference change.

1.3 Modelling Categories of Preference Change

The contributions to this book develop four major types of models of preference change that can be called derivational models (Chapters 2–4), temporal models (Chapters 5–7), consistency-preserving models (Chapters 8–10), and evolutionary models (Chapter 11).

1.3.1 Derivational Preference Change Models

If one kind of preference is linked to another, more basic, kind of preference, then a change in the link between these two preference kinds provides a possible explanation for changes in the non-basic kind of preferences. The most common intuition interprets this relation as a doxastic link, and the resulting change as doxastic preference change. This interpretation lies, for example, at the basis of orthodox decision theory. Savage (1954) proposes that decision problems can be represented by a set of possible consequences $f(s)$, a set of states s of the world and a set of acts f, which take each state of the world to a consequence. The theory connects both to preferences over acts preferences over consequences and beliefs about the states of the world. To this end, desirability of consequences is represented by a real-valued utility function, hence for any consequence $f(s)$, $util(f(s))$ is its utility; the more desirable consequences have higher utility. The agent's beliefs about the state of world are represented by a probability function p on the set of states. These attitudes then determine the agent's preferences over acts: the preferences are represented by her or his expected utility, in the sense that the agent prefers one act to another if the expected utility of the former is larger than that of the latter. The expected utility of an act f is computed as

$$\sum_{s \in S} prob(s) \times util(f(s))$$

from the utility u of its consequences $f(s)$, weighted by the probability $p(s)$ of the state s in which $f(s)$ obtains. Given these dependencies, a change of preferences over acts can be explained as a change in the agent's beliefs about the probabilities of states, given that preferences over consequences remain stable. Standard Savagian decision theory can therefore be applied to preference changes of this form (for an example, see Cyert and DeGroot 1975). Brian Hill's essay in this volume further investigates how the classical decision theoretic framework can be employed for the explanation of preference changes. Following Elster (1982, 1983), he takes Aesop's fable of the fox and the sour grapes as his exemplary scenario of preference change, and offers three analyses of it: (i) In models in terms of pure utility change the fox changes his evaluation of what the grapes would taste like. (ii) In models involving belief revision, at least an external modeller will say that the fox has learnt that the grapes are harder to reach than he thought, changing the overall expected utility. (iii) A third analysis is offered that extends the former two, adding a 'measure of reliability' for the chances of success of an act, in this case, reaching for the grapes.

To take preferences over consequences as basic may be too limiting for many preference change phenomena. Preferences over consequences may themselves be subject to doxastically driven changes, if the preferring agent learns that a consequence has different properties than previously thought. Intuitively, one may think of preferences over such properties as 'values' and of the non-basic ones as preferences over states in which some of these values are realized and others not. Proponents of value atomism (Harman 1967; Quinn 1974; Carlson 1997) defend such a

position – viz. that value has its origin in a few very abstract properties of the world. Pettit (1991) argues that 'choosing on the basis of the properties displayed by the alternatives' captures 'choosing for a reason'. Based on this intuition, one can explain changes in preferences over states as changes in the agent's beliefs about which values these states realize, given stable values.

Van Benthem's contribution surveys dynamic logics of preference change, first for individual agents, and eventually also for groups of agents. He first discusses various formal approaches that start from an ordering of worlds, and derive notions of preference that apply to propositions. Secondly, he offers formal tools to make object or world comparisons on the basis of *criteria*, taking into account the ways in which we apply these criteria, and prioritize between them. Thirdly, he introduces recent developments in the logic of ceteris paribus preferences. On this basis, *dynamic logics* of preference change then describe agents' changing preferences over time, as basic comparison relations for worlds change under model transformations induced by commands, suggestions, or other triggering events that can change preferences. Finally, he discusses logics that intertwine preferences with beliefs. Each has a dynamic aspect, so that we obtain combined dynamic logics for preference change and belief revision, which may be entangled in several ways.

De Jongh's and Liu's contribution develops Benthem's second approach further. They provide a model of changing preferences over objects and show how these preferences can be derived from priorities. Priorities (a concept borrowed from optimality theory) are properties of these objects. For the cases of complete information and (fallible) beliefs, as well as for single and multi-agent cases, they construct different preference logics, some of them extending the standard logic of belief. They then present representation theorems that describe the reasoning valid for preference relations that have been obtained from priorities. Based on these logics, they study preference change with regard to changes of the priority sequence, and changes of beliefs.

Yet another alternative interpretation is found in the concept of a household production function. In this interpretation, the household acquires 'goods' in the market and transforms them through a 'household production function' into 'commodities'. For example, the commodity 'seeing a play' depends on goods like actors, script and theatre, as well as on the consumer's productive input in terms of listening, watching and comprehending (Becker and Chiswick 1966). As another example, Lancaster (1966) suggests 'a glass of orange juice' as a good, from which a consumer with appropriate abilities produces commodities or 'characteristics' like calories and Vitamin C. These commodities or characteristics, rather than the goods, are the arguments of the consumers' utility function. Goods and production abilities are not desired for their own sake, but only as inputs for the production of desired commodities. Thus, a change in preferences over goods can be explained as a change in preference over the production abilities, given stable preferences over commodities. Obviously, changes in production abilities need not be driven by belief changes. For example, one may acquire type-writing skills through sufficient practice, and hence come to prefer type-writing letters over hand-writing them. This

is a preference change that does not seem to be driven by a belief change (although it is accompanied by the acquisition of the belief that one can type faster than before).

Finally, a third interpretation is based on the idea that people adapt their preferences to their abilities and their circumstances, subject to their overall goals. For example, Ng and Wang (2001) suggest that people adjust their attitudes towards income in an optimising fashion, the result being that individuals with low income tend to adopt an attitude with less emphasis on the importance of economic prosperity. Similarly, Welsch (2005) examines a model in which people's preferences adjust to changes in their relative ability to attain various goals. Preference changes are modelled as changes in the configuration of weights attached to these goals. Changes in the individual's opportunity set caused by changes in the attainability coefficients trigger adaptation of the weights attached to the various goals. Hill's suggestion (analysis (iii) in his chapter) that preference change may sometimes be a consequence of changing beliefs about the decision situation (and not about the world), is another example of this category of preference change.

All of these types of derivational preference change can be represented with the help of Richard Jeffrey's decision theory, which we briefly discussed at the end of Section 1.2.3. Jeffrey generalised Savage's framework by replacing the hierarchy of states, acts and consequences in the utility function's arguments with the uniform landscape of propositions. Any proposition could have a utility assigned, and Jeffrey's theory shows how utilities of various propositions can be connected to and restricted by each other. Based on this general framework, Jeffrey (1977) provides a simple model of preference change as the consequence of an agent coming to believe a proposition A to be true. The preferences are represented by a utility function U over propositions. It is defined as the weighted average of the utility u of all the possible worlds w in which the proposition X is true:

$$U(X) = \frac{\sum\limits_{w \in X} u(w) \times P(w|X)}{P(X)}$$

where P is the probability weight (Jeffrey's original notation here is adapted to the discrete case). Now if $\langle u, P \rangle$ represents the preference ordering $\succeq$ that holds before the belief change, and the agent changes her belief $P(A) < 1$ to $P_A(A) = 1$, then $\langle u, P_A \rangle$ represents the changed preference ordering $\succeq_A$ after the belief change. Jeffrey shows that the posterior utility function U_A is related to the prior utility function U as follows:

$$\begin{aligned} U_A(X) &= \frac{\sum\limits_{w \in X} u(w) \times P_A(w|X)}{P_A(X)} \\ &= \frac{P(A)}{P(A \cap X)} \times \frac{\sum\limits_{w \in X \cap A} u(w) \times P(w|A \cap X)}{P(A)} \\ &= U(A \cap X) \end{aligned}$$

One can see clearly how utilities over propositions are derived from utilities over worlds. Crucially, the derivation relation is considerably wider than an instrumental relation between ends and means. Beliefs can influence preferences without relating the relata to some ends towards which these relata contribute. For example, one's preference for winning a trip to Florida in the lottery will crucially depend on one's belief about the weather there during the specified travel time, even though the weather is in no way a means towards winning the trip. Hence, derivational models have a scope beyond models of instrumental relations.

1.3.2 Temporal Preference Change Models

Models involving time preferences analyze preference change on the basis of the temporal occurrence of the preference alone. Time preferences are thus best specified as a relation over pairs of the form (a, t_1) where a is a timeless proposition or sentence, and t_1 some point in time. The particular character of time preferences consists in their dependence on the time factor. Thus it may be the case for instance that an agent consistently holds $(a, t_1) \succeq (b, t_2)$ *and* $(b, t_1) \succeq (a, t_2)$. Insofar as this temporal factor of evaluations can be separated from time-independent factors of evaluations, one speaks of *pure time preferences*.

The standard approach to this issue in economic analysis treats preference as based on value. Value is dealt with in a bifactorial model, in which the value of a future good is assumed to be equal to the product of two factors. One of these factors is a time-independent evaluation of the good in question, i.e. the value of obtaining it immediately. The other factor represents the subject's pure time preferences. It is a function of the length of the delay, and is the same for all types of goods. The most common type of time preference function can be written

$$v(a, t_2) = v(a, t_1) \times (1 - r)^{t_2 - t_1}$$

where r is a discount rate and t_2 a point in time later that t_1. This is the *discounted utility model*, proposed by Samuelson (1937), which still dominates in economic analysis.

There is a wealth of evidence that the discounted utility model does not adequately represent human behaviour. For a simple example, consider a person who prefers one apple today to two apples tomorrow, but yet (today) prefers two apples in 51 days to one apple in 50 days. Although this is a plausible preference pattern, it is incompatible with the discounted utility model. It can however be accounted for in a bifactorial model with a declining discount rate. George Ainslie pointed out that in a single choice between a larger, later and a smaller, sooner reward, inverse proportionality to delay would be described by a plot of value by delay that had a hyperbolic shape. He demonstrated the predicted reversal in pigeons (Ainslie 1974). A discount function with a hyperbolic shape implies a reversal of preference from the larger, later to the smaller, sooner reward for no other reason but that the delays

to the two rewards got shorter. Hyperbolic discounting functions have been widely accepted, at least amongst behavioural economists, as an essentially correct description of people's temporal preferences (Loewenstein et al. 2002). In his contribution to this book, George Ainslie states that the basic hyperbolic shape of discounting is likely to be 'hardwired'. Nevertheless, many think that hyperbolic discounting in humans is in some way 'irrational', or as Ainslie says, 'maladaptive'. Agents afflicted by temporally driven preference reversals experience 'time inconsistencies' that make it hard for them to follow plans they had developed for their own benefit. Thus there is an interesting conflict between temporally driven preference change in accordance with a hyperbolic discount function and strategies to prevent or reverse such preference changes in order to achieve superior results. Spohn's, McClennen's and Ainslie's papers focus on such preventive or reversive strategies, which give rise to preference changes in their own right.

Spohn offers a critique of existing models of rational intertemporal choice under preference change. He devises what he calls a 'global decision model' and argues that this model characterises and generalises all received models of intertemporal choice. He then shows that the global decision model is incomplete, in that it lacks crucial information for a unique prediction or prescription. Different decision rules can be legitimately applied to global models, yielding differing results. Which rule is adequate depends on certain contextual information. Yet as Spohn shows with two examples, this information is not contained in the global decision model, but must instead be taken from somewhere else. Thus, he concludes, current models of intertemporal choice under preference change are incomplete, and fail to account for how agents deal with intertemporal inconsistency.

McClennnen discusses exploitable preference changes. Temporally driven preference changes are exploitable if others can predict an agent's preference reversal, for instance by buying a good from her when she prefers something else, and then later, when she comes to prefer it, sell it to her again at a profit. Hyperbolic discounting seems to lead in many cases to exploitable preference changes. McClennen nevertheless argues that such preference changes need not be ruled out as irrational, because there are rational intertemporal decision rules that hedge against potential exploitations. He argues against the widely accepted *sophisticated choice* rule, with a new argument against the underlying Backward induction principle. Instead he proposes an argument for the *resolute choice* rule, which goes beyond his earlier account (McClennen 1990). Agents who find themselves in a situation in which their preference change may be exploited, and who are aware of this, change their preferences in such a way as to avoid this exploitation. Awareness of possible exploitation thus acts as a determinant on preferences in the direction of time consistency.

Ainslie's contribution can be seen as an additional support of resolute choice. He proposes *recursive self-prediction* as a corrective and stabilizing mechanism arising from the awareness of hyperbolic discounting. Recursive self-prediction allows a notion of *will* that is grown 'from the bottom up', through the selection of increasingly sophisticated processes by elementary motivations. For instance, a dieter faces a tempting food, guesses that she can expect to resist such foods in the future if and only if she resists this particular example, and applies the consequences of this guess

to her current reward contingencies. Her recursive self-prediction thus strengthens her preference to keep the diet, and prevents a preference reversal as the time for potential consumption of such food approaches. In effect she has modified the bargaining game of repeated prisoner's dilemma to describe how a person's successive motivational states relate to each other. The person's will is not presented as a monolithic human faculty, as conceived in the Cartesian tradition, but as something that grows from the person's awareness of how her future motives will be affected by her current choice.

1.3.3 Consistency-Preserving Preference Change Models

Preference change has also been modelled as a consequence of threatening preference inconsistency. As discussed in Section 1.2.4, preference inconsistency arises if the set of preferences violates some of the stipulated preference axioms. To avoid preference inconsistency, agents may have to abandon some of their preferences or add new ones.

Various interpretations exist for this mechanism. According to the theory of cognitive dissonance (Festinger 1957), an individual experiences psychological discomfort when her motivations are inconsistent with one another. In its modern incarnation (see Aronson 1992), the theory argues that an individual's dissonance is particularly acute when this inconsistency reflects on her self-image. Thus, if social status is considered an important aspect of one's self-image, individuals who expend resources in the pursuit of status but fail to attain status experience dissonance. To soften their dissonance individuals may expend greater resources in status seeking, as the positional treadmill approach predicts or, as much of the psychological literature predicts, change their attitudes regarding how status is measured (Oxoby 2004).

The relation of preferences and personal identity or a person's sense of self has also been analysed in an economic context (Akerlof and Kranton 2000; Belk 1988; Frederick 2003). Akerlof and Kranton argue that individuals choose actions and (to some extent) social categories to which they view themselves and others as belonging. In selecting these categories, individuals choose the groups with which they identify. In a similar vein, one can think of the adaptation of attitudes regarding social status as a move towards identifying with various segments of the population (i.e., the underclass or mainstream society). More generally, Akerlof and Kranton propose that identity change may be the result of changing abilities (like the ability to perform or appreciate music vs. the ability to perform sports), and that identity change results in changing one's value profile (cf. Welsch 2005).

Inconsistency may arise not only between preferences, but also between preferences and experienced satisfaction. Conflicts between expectations and experience may lead to cognitive incongruity. Various degrees of incongruity will lead to more or less intensive emotional experiences. In the case of slight incongruity, which only demands assimilative processes, the affective experience is intensified and positively varied. (Ainslie discusses a special case of this with respect to self-prediction in his

chapter.) Unsuccessful as well as some successful attempts to accommodate new information will, though, result in negative experiences. Events that can be adapted to an alternative schema after cognitive processing, that is, occasions of delayed congruity, are generally experienced as positive (see for example, Mandler's 1982 'conflict theory of emotion').

Cognitive incongruity offers an alternative interpretation of how threatening inconsistency can lead to preference change. Cohen and Axelrod (1984) assume that beliefs about the real world are almost always misspecified. Under misspecification, agents will experience 'surprise', as a difference between utility expected from an action and utility experienced after the action. They propose a model of preference change that is in essence a learning process through which agents come to ascribe additional value to means if such means are associated with positive surprises, and come to ascribe less value to a means when it is associated with negative surprises. The model thus shows how agents may come to attribute value to means apart from the instrumental relationship to desired ends, and how these preference orderings of means can change even if the preference ordering over ends remains stable.

Grüne-Yanoff and Hansson propose to model the consistency preserving aspect of preference change after the fashion of belief revision. Theories of belief revision represent processes of changing beliefs that take into account a new piece of information. The logical formalization of these processes has been pursued in philosophy and computer science since the late 1970s. Grüne-Yanoff and Hansson discuss how lessons from belief revision can be applied to modelling preference change. Starting from Hansson's earlier account (Hansson 1995), they argue that while the general input-assimilating framework from belief change can be transferred, several modifications are necessary. The input model has to be complicated with the introduction of a distinction between primary (non-linguistic) and secondary (linguistic) inputs. The method of sentential representation has to be used with somewhat more caution for preferences than for beliefs. Not least, the priority-setting mechanism has to be adjusted, and priority-related information must be included in the inputs.

Rabinowicz critically examines Richard Hare's influential argument for preference utilitarianism, which crucially rests on a model of consistency-driven preference change. Hare suggested that all interpersonal preference comparisons can be reduced to intrapersonal comparisons by asking the agent to form preferences with respect to various hypothetical situations ("what do I prefer for the case in which I were in that person's shoes?") and then balance these preferences against each other. Rabinowicz identifies a gap in Hare's argument, namely that the preferences of Hare's deliberator refer to different hypothetical situations and hence do not enter into conflict. To overcome this difficulty, he considers two different solutions. In one of them, preferences concerning different hypothetical situations are brought into consistency in a way analogous to belief revision, by a process of minimal adjustment. In the other solution, which he calls simultaneous preference extrapolation, each of the input preferences is first universalized and only then the balancing process takes place. The latter proposal differs from Hare's approach in that it introduces moral judgements that are *pro tanto* universal prescriptions, before one arrives at the all-things-considered moral judgment that cannot be overridden.

Luc Bovens discusses Nudge, a new policy style that uses results from behavioral economics and cognitive psychology to affect preferences and choices. Nudge consists in manipulating people's choices in their own interest through arrangements of the choice architecture. A typical example is to induce customers in a self-service cafeteria to choose healthy food by manipulating the order in which the food is presented on the shelves. Nudge seeks to induce people to make better choices, avoid systematic deliberative mistakes and failures of self-control, while respecting their freedom of choice. Bovens argues that Nudge is distinct from other policy instruments such as social advertisement in the way that it seeks to influence preferences, viz. by exploiting patterns of irrationality and circumventing reasons. He investigates to what extent Nudge succeeds in its aims. It may just have local behavioral effects without changing a person's overall preference structure, leading to a fragmented self. It may stand in the way of building moral character, leading to infantilisation. Such cases, he argues, raise questions about the moral permissibility of Nudge-style policies.

Decision theorists have sought to expand decision theoretic frameworks (as discussed in Section 1.3.1) to incorporate consistency preserving preference change. The natural starting point for such an endeavour is Jeffrey's (1977) account of preference change. Jeffrey's model is, however, restricted: It requires an evaluative function u defined over the atoms of the propositional space, viz. possible worlds. Thus for all doxastically changed preference orderings, the preferences over worlds remain identical.

Richard Bradley lifts this restriction in his model of preference kinematics. Expanding on earlier work (Bradley 2005, 2007a, b), he offers a generalization of Jeffrey's Bayesian approach to belief revision, and adds on a preference revision component. In his framework preference change can be described without assuming that fundamental preferences are invariant over persons and time. Desires are expressed in a normalized value function over an algebra of elementary prospects. States of mind are pairs of a probability measure p, standing for the degree of belief, and such a value function v. Preference change is then modelled as an external shock on either beliefs or desires. The dynamics is thus represented by a shift from a state of mind $<p, v>$ to a state $<p', v'>$ caused by a change in p or v. Both belief changes and changes in desire are modelled by extensions of the rules proposed by Jeffrey.

1.3.4 Evolutionary Models of Preference Change

The so-called Indirect Evolutionary Approach (IEA) models the evolution of preferences in a population of agents who rationally choose their strategies to satisfy their preferences (Güth and Yaari 1992; Güth 1995; Huck and Oechssler 1999; Ostrom 2000; Heifetz et al. 2007b). The basic idea is that preferences induce behaviour, behaviour determines 'success', and success regulates the evolution of preferences. What is meant here is *reproductive* success: the ability of a preference

to increase its reproduction, through the behaviour that it induces. In a biological interpretation, this means that the behaviour increases the number of the preference-carrier's offspring, who are genetically endowed with the same preference. In a social interpretation, this means that the behaviour leads to an increased adoption of the preference by others, maybe through learning or imitation.

The mechanism that drives this reproductive advantage is the combined ability of an agent to *commit* to non-equilibrium strategies, and to *signal* this commitment to others. In certain games, such an ability induces opponents to adjust their strategy choices in a way that enhances the fitness of this agent. Consider the following example in Fig. 1.1.

	L	**R**
T	6,2	4,4
B	5,1	2,0

Fig. 1.1 An Inefficient Equilibrium

The strategy T strictly dominates B, and R is a strict best response to T. The unique Nash equilibrium is thus (T, R). However, if player 1 could commit to playing B, and make this commitment known to player 2, then player 2 would respond – in order to maximise her utility – by choosing L. This would lead to result (B, L), a result better for player 1 than the Nash equilibrium (T, R).

But how can player 1 make such a commitment? In IEA, nature makes this commitment for the players, by endowing them with preferences that distort fitness values. Players choose their strategies with the aim of maximising the satisfaction of their preferences over these outcomes, not the fitness outcomes themselves. As IEA shows, having such 'distorted' preferences may enhance fitness results. Take the following example. The left table of Fig. 1.2 is the same game as Fig. 1.1. Payoffs now are interpreted as reproductive fitness results. But 'Nature' distorts player 1's preferences in such a way that strategy B strictly dominates strategy T (leading to the utilities of the right table of Fig. 1.2). Assuming that player 2 knows about player 1's 'distorted' preferences, she will choose L as her rational best reply in the game of Fig. 1.2, leading to outcome (B, L).

A player with 'distorted' preferences obtains a fitness level 5 in this game, while a player with 'undistorted' preferences only obtains a fitness level of 4. 'Distorted'

	L	**R**
T	6,2	4,4
B	5,1	2,0

$\rightarrow$

	L	**R**
T	6,2	4,4
B	8,1	5,0

Fig. 1.2 Preference Distortions

preferences will thus reproduce faster than 'undistorted' preferences, and will not be driven out by any evolutionary process involving payoff-monotonic selection dynamics. 'Distorted' preferences are – in this game and with the given distortion possibilities – *evolutionarily stable*.

Various non-standard preferences have been discussed in this fashion. The idea of indirectness – albeit not in a formal evolutionary framework – was developed by Robert Frank (1987). In his article 'If Homo Economicus Could Choose His Own Utility Function Would He Choose One with a Conscience?' he argues that non-standard preferences are advantageous through their function as commitment devices. Having a conscience, caring about fairness, or experiencing anger may be states that in their direct consequences seem more impedimental than advantageous. Yet they commit agents with those preferences to certain ways of actions – for example, rejecting fraudulent deals, because they are unfair or against one's conscience – hence inducing opponents to actions that lead to more advantageous outcomes. Various authors have used evolutionary game theory to make this idea of indirectness more precise. For restricted sets of preferences and classes of games, Güth and Yaari (1992) show that preferences for reciprocating others' behaviour are evolutionarily stable. Under similar restrictions, others have shown the stability of envious and malevolent preferences (Bolle 2000), altruistic and spiteful preferences (Possajennikov 2000), preferences for fairness (Huck and Oechssler 1999), preferences for relative rather than absolute success (Koçkesen et al. 2000), and social status (Fershtman and Weiss 1998). All of these results are obtained by assuming perfect observability of preferences.

These results have been extended in two directions. Dekel et al. (2007) show that even when allowing for all possible preferences in the population, under perfect observability, efficient, non-equilibrium play is evolutionarily stable in general games. Heifetz et al. (2007a) similarly show that the emergence of 'distorted' preferences is generic, but use a more sophisticated dynamic approach.

Dekel et al. (2007) also show that without observability, the evolutionary stability of 'distorted' preferences breaks down. However, investigating partial observability, Heifetz et al. (2007a) find that inefficient equilibria are destabilized even if a small degree of observability is possible. Güth (1995) and Dekel et al. (2007) obtain similar results.

Evolutionary models contribute to the study of preference change because they provide a model of the context-sensitivity of the frequency with which a certain preference is found in the population. In particular, these models exhibit the sensitivity of preference frequencies to *other* preference frequencies in the population. Preference change is thus presented as a consequence of a changing strategic environment.

In this contribution to this volume, Güth, Kliemt and Napel propose an indirect evolutionary model that investigates the evolution of preferences for trust and trustworthy behaviour. They present a simple trust game where a second mover, the trustee, may have an incentive to cheat a first mover, the trustor. The profitability of the trustor's action depends on the likelihood that the trustee's preferences induce trustworthy behavior. It is assumed that the type composition of the population determines the trustor's beliefs. The trustor decides in advance whether to invest

in the recognition of the trustee's type or not. If she does, then she plays according to posterior beliefs formed in view of the signal she receives. If, however, she does not invest, then play depends on prior beliefs only. It is optimal not to invest if the fraction of trustworthy individuals in the population is very high, or if it is very low: little extra information can be obtained by costly detection activity in either case. Without a risk of detection cheaters fare better than trustworthy individuals, and hence their population share increases. The number of trustworthy individuals will go down all the way to 0 if the initial population share of trustworthy individuals was below the lower bound at which type detection becomes profitable. If on the other hand the initial population share of trustworthy individuals was very high, then it will decrease only until it becomes rational for trustors to invest in obtaining the signal. It turns out that population-dependent parameters can lead to a multiplicity of potentially evolutionarily stable bimorphisms.

1.4 Conclusion

This books presents four fundamentally different types of models of preference change, as outlined above. We believe that this is an example of an area in which methodological pluralism, and in particular a plurality of models, is useful. The reason for this is that preference change is a multifarious topic with many aspects in need of detailed study. Since no sufficiently simple formal model is available that covers all these aspects, we have use for complementary models that elucidate different such aspects. However, this being said, it should be added that the construction of somewhat more comprehensive models that combine some of the features of those presented here would in all probability be a useful addition to the literature.

References

Ainslie, G. W. 1974. Impulse Control in Pigeons. *Journal of the Experimental Analysis of Behavior* 21: 485–489.

Akerlof, G. A. and Kranton, R. E. 2000. Economics and Identity. *Quarterly Journal of Economics* 115: 715–753.

Aronson, E. 1972/2008. *The Social Animal*. 10th edn., New York: Worth/Freeman.

Aronson, E. 1992. The Return of the Repressed: Dissonance Theory Makes a Comeback. *Psychological Inquiry* 3: 303–311.

Barry, H. III, Child, I. L. and Bacon, M. K. 1959. Relation of Child Training to Subsistence Economy. *American Anthropologist* 61: 51–63.

Becker, G. S. 1993. Nobel Lecture: The Economic Way of Looking at Behavior. *The Journal of Political Economy* 101(3): 385–409.

Becker, G. S. 1996. *Accounting for Tastes*. Cambridge, MA: Harvard University Press.

Becker, G. S. and Chiswick, B. 1966. Education and the Distribution of Earnings. *American Economic Review* 56: 358–369.

Becker, G. S. and Michael, R. T. 1973. On the New Theory of Consumer Behavior. *Swedish Journal of Economics* 75: 378–395.

Belk, R. W. 1988. Possessions and the Extended Self. *Journal of Consumer Research* 15(2): 139–168.
Bolle, F. 2000. Is Altruism Evolutionarily Stable? And Envy and Malevolence. *Journal of Economic Behavior and Organization* 42: 131–133.
Bowles, S. 1998. Endogenous Preferences: The Cultural Consequences of Markets and Other Institutions. *Journal of Economic Literature* 36(1): 75–111.
Bradley, R. 2005. Radical Probabilism and Mental Kinematics. *Philosophy of Science* 72: 342–364.
Bradley, R. 2007a. The Kinematics of Belief and Desire. *Synthese* 56(3): 513–535.
Bradley, R. 2007b. A Unified Bayesian Decision Theory. *Theory and Decision* 63(3): 233–263.
Carlson, E. 1997. The Intrinsic Value of Non-Basic States of Affairs. *Philosophical Studies* 85: 95–107.
Chalfant, J. A. and Alston, J. M. 1988. Accounting for Changes in Tastes. *Journal of Political Economy* 96: 390–410.
Cialdini, R. B. and Goldstein, N. J. 2004. Social Influence: Compliance and Conformity. *Annual Review of Psychology* 55: 591–621.
Clark, G. 2007. *A Farewell to Alms*. Princeton, NJ: Princeton University Press.
Cohen, M. D. and Axelrod, R. 1984. Coping with Complexity: The Adaptive. Value of Changing Utility. *American Economic Review* 74(1): 30–42.
Cyert, R. M. and DeGroot, M. H. 1975. Adaptive Utility. In *Adaptive Economic Models*, eds. R. W. Day and T. Groves, 223–246. New York: Academic.
Dekel, E., Ely, J. C., and Yilankaya, O. 2007. Evolution of Preferences. *Review of Economic Studies* 74: 685–704.
Deutsch, M. and Gerard, H. B. 1955. A Study of Normative and Informational Social Influences Upon Individual Judgment. *Journal of Abnormal and Social Psychology* 1: 629–636.
Dreeben, R. 1968. *On What Is Learned in School*. Reading, MA: Addison-Wesley.
Duesenberry, J. S. 1949. *Income, Saving and the Theory of Consumer Behavior*. Cambridge, MA: Harvard University Press.
Edgerton, R. B. 1971. *The Individual in Cultural Adaptation: A Study of Four East African Peoples*. Berkley, CA: University of California Press.
Edvardsson, K., Cantwell, J. and Hansson, S. O. 2009. Self-Defeating Goals. KTH manuscript.
Elster, J. 1982. Sour Grapes: Utilitarianism and the Genesis of Wants. In *Utilitarianism and Beyond*, eds. A. Sen and B. Williams, 219–238. Cambridge/New York: Cambridge University Press.
Elster, J. 1983. *Sour Grapes: Studies in the Subversion of Rationality. Cambridge*. Cambridge University Press/Paris: Maison de Sciences de l'Homme.
Fershtman, C. and Weiss, Y. 1998. Social Rewards, Externalities and Stable Preferences. *Journal of Public Economics* 70: 53–73.
Festinger, L. 1957. *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.
Fisher, F. M. and Shell, K. 1972. *The Economic Theory of Price Indices*. New York: Academic.
Frank, R. H. 1987. If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience. *The American Economic Review* 77(4): 593–604.
Frederick, S. 2003. Time Preferences and Personal Identity. In *Time and Decision*, eds. G. Loewenstein, D. Read and R. Baumeister, 89–113. New York: Russell Sage.
Gaertner, W. 1974. A Dynamic Model of Interdependent Consumer Behaviour. *Zeitschrift für Nationalökonomie* 34: 327–344.
Galbraith, J. K. 1958. *The Affluent Society*. London: Hamish Hamilton.
Galor, O. and Moav, O. 2002. Natural Selection and the Origin of Economic Growth. *Quarterly Journal of Economics* 117: 1133–1191.
Grüne-Yanoff, T. 2004. The Problems of Testing Preference Axioms with Revealed Preference Theory. *Analyse & Kritik* 26(2): 382–397.
Grüne-Yanoff, T. 2008. Why Don't You Want to Be Rich? Preference Explanations on the Basis of Causal Structure. In *Causation and Explanation: Topics in Contemporary Philosophy*, vol. 4, eds. J. Keim Campbell, M. O'Rourke and H. Silverstein, 217–240. Cambridge, MA: MIT Press.
Grüne-Yanoff, T. ed. 2009. *Economic Models – Isolating Tools or Credible Parallel Worlds? Special Issue of Erkenntnis* 70(1).

Grüne-Yanoff, T. and McClennen, E. 2008. Hume's Framework for a Natural History of the Passions. In *David Hume's Political Economy*, eds. C. Wennerlind and M. Schabas, 86–104. London: Routledge.

Güth, W. 1995. An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives. *International Journal of Game Theory* 24: 323–344.

Güth, W. and Yaari, M. E. 1992. An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Behavior in a Simple Strategic Game. In *Explaining Process and Change: Approaches to Evolutionary Economics*, ed. U. Witt, 23–34. Ann Arbor, MI: University of Michigan Press.

Hammond, P. 1976. Changing Tastes and Coherent Dynamic Choice. *Review of Economics Studies* 43: 159–173.

Hansson, S. O. 1995. Changes in Preference. *Theory and Decision* 38: 1–28.

Hansson, S. O. 1996. What Is Ceteris Paribus Preference? *Journal of Philosophical Logic* 25: 307–332.

Hansson, S. O. 2001a. *The Structure of Values and Norms*. Cambridge: Cambridge University Press.

Hansson, S. O. 2001b. Preference Logic. In *Handbook of Philosophical Logic vol 4*, 2nd edn., eds. D. Gabbay and F. Guenthner, 319–393. Dordrecht, The Netherlands: Kluwer.

Hansson, S. O. 2004. Welfare, Justice, and Pareto Efficiency. *Ethical Theory and Moral Practice* 7: 361–380.

Hansson, S. O. and Grüne-Yanoff, T. 2006. Preferences. In *The Stanford Encyclopaedia of Philosophy*, ed. E. N. Zalta. http://plato.stanford.edu/entries/preferences/

Harman, G. 1967. Towards a Theory of Intrinsic Value. *Journal of Philosophy* 64: 792–804.

Harsanyi, J. C. 1953–1954. Welfare Economics of Variable Tastes. *Review of Economic Studies* 21: 204–213.

Heifetz, A., Shannon, C. and Spiegel, Y. 2007a. What to Maximize if You Must. *Journal of Economic Theory* 133(1): 31–57.

Heifetz, A., Shannon, C. and Spiegel, Y. 2007b. The Dynamic Evolution of Preferences. *Economic Theory* 32(2): 251–286.

Henrich, J., Boyd, R., Bowles, S., Gintis, H., Fehr, E., Camerer, C., McElreath, R., Gurven, M., Hill, K., Barr, A., Ensminger, J., Tracer, D., Marlow, F., Patton, J., Alvard, M., Gil-White, F. and Henrich, N. 2005. Economic Man in Cross-Cultural Perspective: Ethnography and Experiments from 15 Small-Scale Societies. *Behavioral and Brain Sciences* 28: 795–855.

Hodgson, G. M. 2008. Review Essay: Prospects for Economic Sociology. *Philosophy of the Social Sciences* 38: 133–149.

Holbrook, M. B. and Schindler, R. M. 1989. Some Exploratory Findings on the Development of Musical Tastes. *Journal of Consumer Research* 16: 119–124.

Holbrook, M. B. and Schindler, R. M. 1994. Age, Sex, and Attitude Toward the Past as Predictors of Consumers' Aesthetic Tastes for Cultural Products. *Journal of Marketing Research* 31: 412–422.

Holbrook, M. B. and Schindler, R. M. 1996. Market Segmentation Based on Age and Attitude Toward the Past: Concepts, Methods, and Findings Concerning Nostalgic Influences on Customer Tastes. *Journal of Business Research* 37: 27–39.

Huck, S. and Oechssler, J. 1999. The Indirect Evolutionary Approach to Explaining Fair Allocations. *Games and Economic Behavior* 28: 13–24.

Jeffrey, R. C. 1977. A Note on the Kinematics of Preference. *Erkenntnis* 11: 135–141.

Jeffrey, R. C. 1983. *The Logic of Decision*. Chicago, IL: University of Chicago Press.

Kahneman, D., Slovic, P. and Tversky, A. eds.. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kahneman, D., Knetsch, J. L. and Thaler, R. H. 1991. Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *The Journal of Economic Perspectives* 5(1): 193–206.

Kapteyn, A. and Wansbeek, T. J. 1982. Empirical Evidence on Preference Formation. *Journal of Economic Psychology* 2: 137–154.

Koçkesen, L., Ok, E. A. and Sethi, R. 2000. The Strategic Advantage of Negatively Interdependent Preferences. *Journal of Economic Theory* 92(2): 274–299.

Koopmans, T. C. 1957. *Three Essays on the State of Economic Science*. New York: McGraw-Hill.

Krelle, W. 1973. Dynamics of the Utility Function. In *Carl Menger and the Austrian School of Economics*, eds. J. R. Hicks and W. Weber, 92–128. Oxford: Oxford University Press.

Laibson, D. 1997. Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics* 62: 443–477.

Lancaster, K. 1966. A New Approach to Consumer Theory. *Journal of Political Economy* 74: 132–157.

Landsburg, S. E. 1981. Taste Change in the United Kingdom 1900–1955. *The Journal of Political Economy* 89(1): 92–104.

Leibenstein, H. 1950. Bandwagon, Snob, and Veblen Effects in the Theory of Consumer's Demand. *Quarterly Journal of Economics* 64: 183–207.

Loewenstein, G. 1996. Out of Control: Visceral Influences on Behavior. *Organizational Behavior and Human Decision Processes* 65: 272–292.

Loewenstein, G. 2000. Emotions in Economic Theory and Economic Behavior. *American Economic Review: Papers and Proceedings* 90: 426–432.

Loewenstein, G. and Angner, E. 2003. Predicting and Indulging Changing Preferences. In *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, eds. G. Loewenstein, D. Read and R. Baumeister, 351–391. New York: Russell Sage.

Loewenstein, G. and Schkade, D. 1999. Wouldn't It Be Nice? Predicting Future Feelings. In *Well-Being: The Foundations of Hedonic Psychology*, eds. D. Kahneman, E. Diener and N. Schwarz, 85–105. New York: Russell Sage.

Loewenstein, G. Read, D. and Baumeister, R. eds. 2002. *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*. New York: Russell Sage Foundation Press.

Mandler, G. 1982. The Structure of Value: Accounting for Taste. In *Affect and Cognition - The Seventeenth Annual Carnegie Symposium on Cognition*, eds. Clark and Fiske, 3–36. London/Hillsdale, NJ: Erlbaum.

Marcuse, H. 1964. *One-Dimensional Man*. London: Abacus.

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

Mill, J. S. 1844. On the Definition of Political Economy; and on the Method of Investigation Proper to It. In *In Essays on Some Unsettled Questions of Political Economy*. Reprinted in Collected Works of John Stuart Mill, ed. J. M. Robson, Vol. 4, 309–339. Toronto: University of Toronto Press/London: Routledge & Kegan Paul, 1963–1991.

Ng, Y. K. and Wang, J. 2001. Attitude Choice, Economic Change, and Welfare. *Journal of Economic Behavior and Organization* 45: 279–291.

Nisbett, R. and Ross, L. 1980. *Human Inference: Strategies and Shortcomings of Human Judgement*. Englewood Cliffs, NJ: Prentice-Hall.

Ostrom, E. 2000. Collective Action and the Evolution of Social Norms. *Journal of Economic Perspectives* 14(3): 137–158.

Oxoby, R. J. 2004. Cognitive Dissonance, Status and Growth of the Underclass. *Economic Journal* 114: 727–749.

Parsons, T. 1934. Some Reflections on "The Nature and Significance of Economics". *Quarterly Journal of Economics* 48(3): 511–545.

Parsons, T. 1937. *The Structure of Social Action*. 2 vols. New York: McGraw-Hill.

Parsons, T. 1970. On Building Social Systems Theory: A Personal History. *Daedalus* 99: 826–881.

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

Pettit, P. 1991. Decision Theory and Folk Psychology. Reprinted in *Rules, Reasons, and Norms: Selected Essays*, 192–221. Oxford: Oxford University Press, 2002.

Polanyi, K. 1944. *The Great Transformation*. Boston, MA: Beacon.

Pollak, R. A. 1968. Consistent Planning. *Review of Economic Studies* 35: 201–208.

Pollak, R. A. 1976a. Interdependent Preferences. *American Economic Review* 66: 309–320.

Pollak, R. A. 1976b. Habit Formation and Long-Run Utility Functions. *Journal of Economic Theory* 13: 298–318.
Pollak, R. A. 1977. Price Dependent Preferences. *American Economic Review* 67(2): 64–75.
Pollak, R. A. 1978. Endogenous Tastes in Demand and Welfare Analysis. *American Economic Review* 68(2): 374–379.
Possajennikov, A. 2000. On the Evolutionary Stability of Altruistic and Spiteful Preferences. *Journal of Economic Behavior and Organization* 42(1): 125–129.
Potter, D. 1954. *People of Plenty: Economic Abundance and the American Character*. Chicago, IL: University of Chicago Press.
Quinn, W. S. 1974. Theories of Intrinsic Value. *American Philosophical Quarterly* 11: 123–132.
Rescher, N. 1967. Semantic Foundations for the Logic of Preference. In *The Logic of Decision and Action*, ed. N. Rescher, 37–79. Pittsburgh, PA: University of Pittsburgh Press.
Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. 1st edn., London: Macmillan.
Samuelson, P. A. 1937. A Note on Measurement of Utility. *Review of Economic Studies* 4: 155–161.
Savage, L. J. 1954. *The Foundations of Statistics*. New York: Wiley.
Schindler, R. M. and Holbrook, M. B. 1993. Critical Periods in the Development of Men's and Women's Tastes in Personal Appearance. *Psychology & Marketing* 10: 549–564.
Schulkin, J. 1991. *Sodium Hunger*. Cambridge: Cambridge University Press.
Schumpeter, J. 1942. *Capitalism, Socialism and Democracy*. New York: Harper & Row.
Schwarz, N. and Strack, F. 1991. Context Effects in Attitude Surveys: Applying Cognitive Theory to Social Research. *In European Review of Social Psychology* 2: 31–50.
Smith, A. 1776. *The Wealth of Nations*. Reprint. New York: Random House.
Stigler, G. J. and Becker, G. S. 1977. De gustibus non est disputandum. *American Economic Review* 67: 76–90.
Strotz, R. H. 1956. Myopia and Inconsistency in Dynamic Utility Maximization. *Review of Economic Studies* 23: 165–180.
Tourangeau, R. 1992. Context effects on attitude responses: The role of retrieval and necessary structures. In context effects in social and psychological research, ed. N. Schwarz and S. Sudman, 35–47. New York: Springer.
Veblen, T. 1899. *Theory of the Leisure Class*. New York: Macmillan.
von Weizsäcker, C. C. 1971. Notes on Endogenous Changes of Tastes. *Journal of Economic Theory* 3: 345–371.
von Wright, G. H. 1972. The Logic of Preference Reconsidered. *Theory and Decision* 3: 140–169.
Weisbrod, B. 1977. Comparing Utility Functions in Efficiency Terms, or, What Kind of Utility Functions Do We Want. *American Economic Review* 67: 991–995.
Welsch, H. 2005. Adaptation of Tastes to Constraints. *Theory and Decision* 57(4): 379–395.
Winston, G. C. 1980. Addiction and Backsliding: A Theory of Compulsive Consumption. *Journal of Economic Behavior and Organization* 1: 295–324.
Yaari, M. E. 1977. Endogenous Changes in Tastes: A Philosophical Discussion. *Erkenntnis* 11: 157–196.

Chapter 2
Three Analyses of Sour Grapes

Brian Hill

Abstract The phenomenon of adaptive preferences – sometimes also known under the name of *sour grapes* – has long caused a stir in Social Theory, mainly because of its importance in the debate over utilitarianism. The question of preference change has been considered by decision theorists and, more recently, logicians. The former phenomenon seems a natural candidate for application of the latter theories. The fundamental question of sour grapes is: what is it that changes – the agent's beliefs or his utilities? The aim of this paper is to consider the replies that decision theorists and logicians can offer to this question. Besides the interest of the phenomenon as a case study for theories of change, it raises two general points. Firstly, besides a belief change and a utility change, there is a third possibility for the source of a given change in preferences: a change in the decision-maker's perception of the choice he is faced with. Secondly, traditional methods for eliciting beliefs and utilities do not function well in cases where several situations are involved and the relations between the agent's attitudes in the different situations are at issue. An elicitation method is sketched which purports to deal more adequately with such cases. Although based on independent motivations, it provides another argument for the importance of taking into account how the decision-maker perceives the choice he is faced with.

2.1 What is Sour Grapes?

2.1.1 The Phenomenon and the Challenge

In La Fontaine's fable, the fox approaches a tree, attempts to reach the grapes, and, realising that he cannot, turns away, saying to himself that they were sour.

This phenomenon, which goes under the name of adaptive preferences or sour grapes, has long caused a stir in Social Theory, mainly because of its importance

B. Hill
GREGHEC, HEC Paris and IHPST. 1 rue de la Libération, 78351 Jouy-en-Josas, France
e-mail: hill@hec.fr.

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

in the debate over utilitarianism (Elster 1983; Sen and Williams 1982). After all, if people's desires change so whimsically, how could any aggregate of these be a proper guide for social choice? But the phenomenon also involves many themes of recent and ancient interest to decision theorists and logicians, such as choice and preference change. The aim of this paper is to see what understanding these theories can offer of the sour grapes phenomenon. More important than the particular challenge posed by this specific phenomenon will be general issues it raises.

Throughout this paper, we will concentrate on the example in La Fontaine's fable. It involves two acts – one being the first attempt at getting the grapes, the second being the act of walking away – between which the fox has changed his mind. The challenge of sour grapes is to understand the relationship between the pair of acts and the potentially changing attitudes on which they are based.

An appropriate framework for couching discussion of attitudes, their role in action, and their changes is classical Savagean decision theory. Faced with a choice among several options, a (rational) agent establishes what he considers to be the possible consequences of the options and identifies the facts about the world which affect the realisation of these consequences given that the option is chosen; subsequently, he chooses the option which, given what he believes about the relevant properties of the world, yields on average the most attractive consequences. Formally, the decision problem is represented as a set of possible consequences, a set of states of the world and a set of acts – functions taking each state of the world to a consequence – which are the objects over which the agent is to choose. The agent is taken to have a preference relation over the acts, which determines his choice. We shall call the tuple consisting of a set of states, a set of consequences and a preference relation over the acts taking these states to these consequences a *decision situation*. According to classical decision theory, the attractiveness of the consequences in a decision situation for the agent are represented by a real-valued *utility function*: the more attractive consequences have higher utility. In a similar way, the agent's beliefs about the state of world are represented by a *probability function* on the set of states. These attitudes determine his preferences over acts: the preferences are represented by his *expected utility*, in the sense that he prefers one act to another if the expected utility of the former is larger than that of the latter. Formally: for a probability p and a utility u, the expected utility of an act f is $\sum_{s \in S} p(s).u(f(s))$. The choice of the expected utility framework is not intended to imply any particular commitment to this fashion of theorising; rather, its purpose is to facilitate precise discussion. Indeed, although it shall be assumed that the agent is an expected-utility maximiser, rather than a maximiser with respect to some other more complicated non-expected utility decision rule, many of the points hold for other models of decision (see below for remarks concerning the relationship to Jeffrey's decision-theoretic framework, which is preferred by many philosophers).

To reformulate the sour grapes story in decision-theoretic terms, the fox's preferences for different available acts (attempting to grasp the grapes, walking away) has changed between his first attempt to get the grapes and his walking away. Assuming him to be minimally rational, this means that the expected utilities of these acts must have changed. However, this change is not in itself the purportedly interesting

property of sour grapes. Expected utility changes are widespread, and can occur for two reasons: on the one hand, such a change can result from a change in the agent's beliefs about the world, on the other hand, it could also be a consequence of change in his utility. For the use to which sour grapes is put in the debate over utilitarianism (Elster 1983; Sen and Williams 1982), it is crucial that they are not just cases of belief change, but cases of utility change.[1] If there is utility change, so the argument against utilitarianism goes, then the utilities of different members of a society may not be stable; however, utilitarianism relies on these utilities, so if they are unstable, the utilitarianist position, at least as it is traditionally stated, is weakened. For this argument to be valid, it needs to be determined that it is indeed utilities rather than beliefs which change in an apparent case of sour grapes.

These considerations indicate the importance of the phenomenon of sour grapes for decision theory. Sour grapes poses the problem of the identity, stability and variability of the central notions of decision theory – utility, belief, expected utility. The basic question of sour grapes is: what changes? Beyond the preferences on acts (expected utilities), is there a change in utilities, or just a change in beliefs? From the point of view of a decision-theorist or a logician, this question constitutes the principal challenge posed by the phenomenon of sour grapes.

One can conceive of two methods for answering this question. The *direct* method relies on decision-theoretic machinery: one uses classical techniques to elicit the beliefs and utilities in the appropriate situations and then compares them to see which have changed and how. This method is not open to theorists who do not have a way of eliciting the agent's attitudes on the basis of his preferences; it can be employed by decision-theorists, for example, but not by logicians. An alternative, *indirect* method consists of finding properties of the change involved in the sour grapes story which are *only* possessed by belief changes or utility changes. To take a very simple example, suppose that, according to a certain theory of attitude change, changes in utility but not changes in belief can cause preference reversals of a certain type; if sour grape phenomena involve preference reversals of this type, then one can conclude that, according to these theories, sour grapes involves utility change. To apply this method, one needs firstly to identify some properties of the sour grapes phenomena, and then consider how different analyses of the change can account for these properties. In Section 2.1.2, some noteworthy properties will be identified; in Section 2.2, the possible analyses of the phenomenon will be explored. A decisive conclusion will not be reached; in Section 2.3, the direct method will be considered.

Remark 1 As indicated above, the difference between utility and expected utility turns out be very important to the understanding of the phenomenon of sour grapes. Given the preceding considerations, the important distinction is the following: utility is *pure* insofar as the calculation of the utility of a consequence does not depend on the beliefs of the agent, whereas expected utility is *mixed*, insofar as it depends not only on the agent's utilities for consequences, but also on his beliefs. However, this

[1] See Elster (1983, pp 112–114) for more details on why learning (a paradigmatic form of belief change) is not necessarily a problem for utilitarianism.

distinction is not always present in other frameworks, and, even where it can be drawn, it is often not explicitly recognised.[2]

First of all, the decision theory proposed by Jeffrey (1972), as well as the causal decision theories inspired by it (for example Joyce 1999), does not accept a distinction of this sort.[3] Whereas Savage distinguishes between the objects of pure utility (consequences), the objects of choice (acts), which have mixed utility, and the things over which the agent has no power (states of the world), Jeffrey employs a single ontology, consisting of "propositions" (sometimes also known as "prospects"). Each proposition has a probability and a utility (or "desirability") value, the latter being related to the probabilities and utilities of other propositions according to the following conditional expected utility formula: for incompatible propositions X and Y, $u(X \vee Y) = u(X).p(X/X \vee Y) + u(Y).p(Y/X \vee Y)$. So everything – from the choice of trying to get the grapes to the prospect of obtaining the grapes to the fact that they are 2 m off the ground – has a utility, and each of these utilities are mixed in the sense that they can be calculated from the utilities of other propositions. (Because of the technical conditions required, every proposition can be written as the disjunction of two incompatible non trivial propositions.) The general question posed by sour grapes – what changes: beliefs or (just) utilities? – makes no sense in this framework, for there is, in general, no change of utilities without a change in beliefs.[4]

Nevertheless, it does make sense to ask about the changes of utilities for specific propositions, or changes of beliefs in particular propositions. In the case of the fox, is it the utilities of the grapes which have changed, or the beliefs about their position? In the first case, there are beliefs which change (for example, beliefs as to whether he will get the grapes), in the second case, there are utilities which change (for example, the utility of the position of the grapes), but these do not seem pertinent for the problem posed by sour grapes. A simple way to understand the discussion of the majority of this paper, which will employ the Savage framework, would be to "embed" the Savage-style decision situations described above into the Jeffrey framework, and understand talk of 'utility change' and 'belief change' as referring to those utilities and beliefs which are present in the Savage framework (i.e. utilities for Savage consequences, and beliefs in Savage states). In this way, a meaningful distinction between pure and mixed utility change could be drawn in the Jeffrey framework.

[2] In cases where theorists of 'preference change' do not draw a difference between utility and preference or expected utility, their theories shall, insofar as is possible, be interpreted as theories of utility change rather than expected utility change; this interpretation, although debatable, at least allows us to consider what their theories can bring to the problem of sour grapes.

[3] Below, when treating alternatives to Savage, we will restrict attention to Jeffrey's evidential decision theory and not consider the causal decision-theoretic variant in any detail. Although there may be differences in the details and debate about which of the two is more naturally applicable to particular cases, the general points made here apply to both sorts of theory. Indeed, our intention is to avoid such debates and details, which are not central to the problem in hand.

[4] This is an immediate consequence of the fact that, in Jeffrey's decision theory, desirabilities determine probabilities (Jeffrey 1972, §5.9).

There appears to be no clear consensus as to what such an embedding should be.[5] Although the details are not important for the purposes of this paper, we shall take the following "brute-force" embedding.[6] For Savage states $s_1, s_2 \ldots$, let there be propositions $S_1, S_2 \ldots$ stating that the appropriate state is realised, and for Savage consequences $c_1, \ldots$, let there be propositions $C_1, \ldots$ stating that the appropriate consequence is realised (note that the S_i and the C_j are each sets of mutually incompatible propositions partitioning logical space). Let F be the proposition that act f is carried out. It follows from Jeffrey's conditional expected utility formula that $u(F) = \sum_{S_i} \sum_{C_j} u(C_j \wedge F).p(C_j/F \wedge S_i).p(S_i/F)$. If one assumes that probabilities of states are act-independent ($p(S_i/F) = p(S_i)$), that utilities of consequences are act-independent ($u(C_j \wedge F) = u(C_j)$), and that the probability of a consequence given an act and a state is one if the act sends that state to that consequence and zero otherwise ($p(C_j/F \wedge S_i) = 1$ if $f(s_i) = c_j$ and 0 otherwise), then one obtains the traditional Savage expected utility formula. Whatever is said about the Savage framework holds for the aforementioned propositions under the assumptions just stated; we shall consider these propositions and these assumptions to constitute the embedding.

The introduction of the distinction between pure and mixed utilities also requires a re-appraisal of several purported theories of "utility change" or "preference change", insofar as they can be applied to sour grapes. There are theories, both in the decision-theoretic and logical literature, which model "utility" or "preferences" as depending on beliefs and which account for utility change in terms of changes of beliefs: Cyert and DeGroot (1975); de Jongh and Liu (2006) are examples. They should be considered as models of change in *expected utility* (or, at least, change in mixed utility), and not of change in (pure) utility. If the fox turns away, as in La Fontaine's version of the fable, saying to himself that the grapes are sour, this may be understood as a change in belief about the taste of the grapes. The utilities for the grapes are thus to be understood as *mixed* utilities, which are a function of the (pure) utility of grape–properties and of the fox's beliefs as to whether these grapes have those properties have changed. Sour grapes will thus be analysed as a change in the fox's beliefs about whether the grapes have the grape–properties in question; under this analysis, the *mixed* utilities have changed, but the *pure* utilities have remained constant (he still enjoys the grape–properties as before). As noted above, the

[5] Normally in the Jeffrey framework, one explicitly does away *either* with states of the world *or* with consequences when considering the utility of an act. In Jeffrey's introduction to his theory (1972, Ch. 1), he does away with consequences, and considers explicitly only states of the world (which he calls "conditions"): using the terminology introduced in the text, $u(F) = \sum_{S_i} u(S_i \wedge F)p(S_i/F)$. Here the consequences are not explicitly represented but replaced by the conjunction of the realisation of the state and the act. By contrast, others (for example Levi (2000)) remove the states and leave the consequences: this yields $u(F) = \sum_{C_j} u(C_j \wedge F)p(C_j/F)$. Here the states are not explicitly represented, but are tied up in the conditional probabilities. Both of these can be derived from the general formula stated in the text: the former by summing over the consequences, the latter by summing over the states. See also Jeffrey (1972, §10.4).

[6] For a more sophisticated account of the relationship between the two frameworks, see Bradley (2007b).

intuition that the pure utilities as well as the mixed utilities change in sour grapes is important for the debate on utilitarianism (Elster 1983, p123); these theories deny this intuition in their analysis of the sour grape phenomenon.

2.1.2 Some Properties of Sour Grapes

Here are some observations concerning sour grapes inspired from Elster's classic discussion of the phenomenon.

The nature of the change Elster (1983, Ch III) emphasises that the change in preferences is of a causal nature, and may not be intentional on the part of the agent. Whatever the fox's opinions on the change, it was *caused* by his experience of the first attempt at obtaining the grapes; he *did not decide* to change his preferences in the face of this experience.

The source of the change Regarding the source of such changes, as Elster (1983, pp121–122) points out, one might draw a distinction between those caused by changes in the world – the "state-dependence" of preferences – and those caused by changes in the options open – "possibility dependence". Almost immediately he qualifies the distinction by noting that, given the possible interdependence of *states* and *options*, and the difficulty in getting a clear separation of the two notions, it may be practically impossible to apply this distinction correctly in practice. One might expect that a proper account of sour grapes take account of this distinction and its instability (or, if you prefer, flexibility).

To avoid confusion, it should be emphasised that the sort of dependence to which Elster seems to be alluding, especially in the case of state-dependence, is *diachronic* dependence. The *previous* states, choices and so on have an influence on the *current* utility. This notion is thus crucially different from the "state-dependence of utility" studied by decision theorists such as Karni and Drèze. A state-dependent utility is a utility that is not only a function of the consequence, but equally of the state which the act will "take" to that consequence. (The expected utility is calculated by the formula $\sum_{s \in S} p(s)u(s, f(s))$, with u having two arguments.) Note that the dependence in this case is entirely *synchronic*: in a single decision situation, at a single instant, the utility depends on the states. Although Elster only claims diachronic state-dependence for sour grapes, the synchronic or decision-theoretic sense is mentioned because it will be relevant below.

The permanence of the change Elster (1983, pp112–114) takes pains to emphasise that the adaption of preferences in situations involving sour grapes is in principle *reversible*, after a further change in the situation. In fact, intuitions regarding the question of reversibility or permanence of the change differ depending on the time-scale involved. In general, there are three basic intuitions. The first dictates that the fox does not instantly and immediately change his mind about the grapes, and so would take them if the possibility arose immediately after his exclamation that they

were no longer desirable. According to such an intuition, reversibility of the purported change is very plausible at moments close to the situation in question. The second intuition arises from the idea that it does seem possible, over a longer period of time, and perhaps through force of habit, for the fox to *actually acquire the sort of attitudes (regarding the grapes) he claims to have*, so that he would *not* take the grapes if offered. Indeed, many pertinent examples of sour grapes generally involve an extended time span over which individuals' attitudes seem to change "for good"; such is the case for the change of preferences for city or countryside life considered by Elster (1983, p112 *sq*.). A final intuition, that which is expressed by Elster, dictates that even this long-term change can be reversed by a change in situation: given the correct situation, the fox would once again act in accordance with a preference for grapes. In Elster's example, someone who moves to the city may acquire a taste for city life, which may be reversed if he moves back to the countryside for a considerable period. A full analysis of sour grapes should be able to account for these factors.

2.2 Three Analyses

This section contains three analyses of sour grapes. In fact, to the extent that they leave precisely specified blanks to be filled by particular mechanisms, they are better described as analysis schemata. For each analysis, its capacity to account for the properties of sour grapes discussed above will be considered. This will not only allow us to ascertain whether the indirect method described in Section 2.1.1 can yield a reply to the question of sour grapes, but it will highlight some challenges for current theories of attitude change.

Throughout the section, the standard sour grape story introduced in Section 2.1 will be the principal object of consideration. Recall that this story involves two decision situations: the situation before the fox's first attempt at getting the grapes (which shall be called 'the first situation'); the situation after this attempt, and in which he takes the decision to try again or to give in (the 'second situation'). These situations seem to share the same set of pertinent states of the world, where in each state factors such as the position of the grapes and the height of the tree are determined. They also share the same set of consequences, which contains two elements – obtaining the grapes and not obtaining the grapes. It follows that the two situations share the same set of acts.

2.2.1 Pure Utility Change

The analysis The simplest analysis of sour grapes takes it at face value: the difference between the two situations is indeed a change in (pure) utilities. In such

a model the fox's utility function is different in the second situation with respect to the first, and this explains his decision not to pursue his attempt to obtain the grapes.

Writing this formally, let the initial preferences of the fox be determined by probability p_1 and utility u_1, so that the expected utility of an act f is $\sum_{s\in S} p_1(s)u_1(f(s))$. Then, according to this model of the change, the probability in the second situation will also be p_1, but the utility will now be u_2. The preferences of the fox (over actions) will thus be represented by $\sum_{s\in S} p_1(s)u_2(f(s))$. A full model results when one adds an account of the change from u_1 to u_2; one might expect theories of utility change for example to provide such accounts.

Properties of the analysis

What has changed Concerning the question of what has changed, it is the (pure) utility which is taken to change in this situation. Indeed, the interpretation of this situation as a utility change is unavoidable, in the following sense: a given change in utility will result in a change in preferences which is such that it there is *no* change in beliefs alone which could have produced this change in preferences. That is to say, one could not even rewrite the representation in the second situation $\sum_{s\in S} p_1(s)u_2(f(s))$ *as if* it consisted in a change of belief with a fixed utility (i.e. in the form $\sum_{s\in S} p'(s)u_1(f(s))$), because for these two sums to represent the same preferences, the probabilities and utilities must be the same ($p_1 = p'$ and $u_2 = u_1$).[7] One can thus conclude that this model of sour grapes essentially involves a change in utility.

The source of the change The source of the utility change is the fox's choice in the first situation and his experience following it: in this sense, the change is state-dependent. On the other hand, the same options (acts) are available in the first and the second situation, so the change cannot be thought of as possibility-dependent. Indeed, many of the current theories of preference or utility change keep the same options (i.e. possibilities) but alter the preferences on them[8]; to this extent, they could only be understood in terms of state-dependent change. As such, they fail to account for the tight relationship between state- and possibility-dependent change.

The nature of the change Theories of utility change are often motivated by examples involving changes in the face of statements specifying particular preference

[7] Suppose not, and, supposing appropriate calibration of the utilities, let p' be a probability such that $\sum_{s\in S} p_1(s).u_2(f(s)) = \sum_{s\in S} p'(s).u_1(f(s))$ for all acts f. Given an appropriately rich set of acts, as implied by the Savage axioms, for example, it follows that $p_1(s).u_2(c) = p'(s).u_1(c)$ for all states s and consequences c. However, given that s and c are independent variables, this implies that $p' = p_1$ and $u_1 = u_2$.

[8] Cyert and DeGroot (1975), van Benthem and Liu (2007), Bradley (2007a) and the revision and contraction operations in Hansson (1995) are examples.

relations to be accepted,[9] and it is not evident how to translate the changes in the world involved in this example – the fox's experience of his first attempt at getting the grapes, say – in terms of such statements. This just flags a general and important issue for theories of change: how is the trigger for a given change to be represented in the model? This question will always need to be posed when, as is the case here, theories are motivated with examples of *intentional* preference change (considered to be the preference analogues of apparently intentional processes such as learning from observation or accepting an announcement): something should be said about how one can apply the same methods for non-intentional changes (as noted in Section 2.1.2, in the case of sour grapes, the fox's experience *causes* the change, he does not *decide* to change his utility after the experience).

The permanence of the change The analysis does seem to account for the case where the subject actually acquires the preference in the long term, modelling it as a straight utility change. However, the modalities of the change seem to have been reversed: an adaptation of the preferences over a long period is captured here by a sudden revision of the utility at a particular moment. Furthermore, it is not certain to what extent the analysis can account for the short- or long-term reversibility of the change, because there is no guarantee that the utility change it proposes is reversible. As for the case of the analysis in terms of belief change proposed below (Section 2.2.2), the capacity to account for such phenomena may depend on the particular theory of utility change adopted.

Many theories of pure utility change are modelled on, or related to, particular theories of belief change. The theory of Cyert and DeGroot (1975) is not strictly speaking a theory of pure utility change but of mixed utility change, insofar as utilities depend on beliefs and the change in utility arises from a change in beliefs, notably by Bayesian conditionalisation. *Even if* it were possible to understand this as a theory of pure utility change, it would inherit the reversibility properties of Bayesian conditionalisation, and notably the fact that this sort of change is generally irreversible – information is lost in the change. Similarly, approaches modelled on public announcement logics also tend to yield change operators which are irreversible (the update and upgrade operators in van Benthem and Liu [2007] seem to be examples).

On the other hand, some theories which may be understood as theories of utility change are inspired by the literature on belief revision: such is the theory of Hansson (1995). They have two operations: revision, which establishes an order on the utilities of various consequences, and contraction, whereby one revokes a particular order on the utilities of consequences. An application of the former operation followed by an application of the latter (with respect to the same order on the same consequences) yields the original utility: in this sense, there is reversibility. Similarly, the theory of Bradley (2007a) according to which utility values are re-allocated

[9] In Hansson (1995), the agent "learns" that a certain outcome has a certain desirability, and alters his preferences accordingly; in van Benthem and Liu (2007), an agent is told to prefer a certain outcome, and alters his preferences accordingly.

on events in a partition, is reversible: it suffices to apply a change operation allocating the original utility values to the elements of the partition to return to the intial utility function.

Although technically, so to speak, these theories give reversibility, the sorts of concerns mooted above regarding the representation of the trigger for change apply here: if, as the fox walks away, he sees a ladder and uses it to get the grapes, is this correctly understood as a retraction of the original utility change whereby his utility for the grapes decreased (as in the case of Hansson's theory) or as the demand to acquire utilities for grapes which happen to be those he had initially (as in the case of Bradley's theory)? Even if this is a correct analysis of the case of short-term reversibility, will the same mechanisms be at work in the long-term case (when he reverts backs to a high utility for grapes after years of having a low utility for grapes)? A final, general worry concerns the intuitive adequacy of any analysis of short-term reversibility of preferences in terms of a succession of utility changes: is it really plausible that the fox's utility undergoes two abrupt changes in such a short period of time? The reversibility and permanence phenomena pose interesting challenges for anyone seeking to take up and defend this analysis, and it is not clear that they are fully met by current models of utility change.

This first analysis of sour grapes is firmly embedded in a developing theory of utility change. At this stage, all that can be noted is the difficulties which should overcome: it does not seem to capture the subtle relationship between the state- and possibility-dependance of the change; careful interpretation of the model is required to make sure it can cope with the non-intentional nature of the change; and there are doubts regarding its capacity to account for the permanence and reversibility properties of the change. It is debatable how many of these difficulties support the conclusion that sour grapes is not a case of utility change and how many are to be seen as concerns with the adequacy of current models of utility change.

Nevertheless, even under this meagre construal, some may find the analysis in terms of utility change inadequate. There is an intuition, which seems incompatible with this sort of analysis, that the fox's utilities do not *really* change, at least not immediately after his failed attempt at getting the grapes. The second analysis of sour grapes takes this as its guiding intuition.

2.2.2 Self-justification

The analysis An important intuition about the sour grapes phenomenon is that it does not involve so much the *action* of the agent (at the moment of the sour grapes phenomenon) as the *way he justifies* or *rationalises* that action (to or for himself). The fox walks away from the grapes in any case; it is the reason he gives himself for walking away that is at issue. Under this interpretation of the phenomenon, although it does not (directly) affect concurrent behaviour, the rationalisation he constructs for himself will affect the utilities and the beliefs *he sees himself as having*.

In this analysis, crucial use is made of the distinction between the point of view of the modelee – the agent – and that of the modeler – the decision theorist. The fact that the modeler elicits certain probability and utility functions representing the beliefs and utilities of the agent does not imply that the agent himself will recognise these as his beliefs or utilities. This allows one to distinguish between an *internal* model, representing the utilities and the beliefs the fox sees himself as having, and the representation of a competent external observer. According to the analysis of sour grapes as simple self-justification, the change in the expected utility is properly thought of as a revision of beliefs with information learnt during the first attempt at getting the grapes: he learns that the grapes are more difficult to obtain than previously thought. However, as opposed to the case discussed in Section 2.2.1, a change in beliefs always produces a change in preferences (expected utility) which can also be produced by a change in utilities (with constant beliefs). This is the change that the fox considers to have occurred: he represents the change to himself as a change in degree to which he values the grapes, that is, as a change in utilities.

As in the previous example, let the initial preferences of the fox be determined by probability p_1 and utility u_1, so that the expected utility of an action f is $\sum_{s\in S} p_1(s)u_1(f(s))$. Furthermore, assume that $p_1(s) \neq 0$ for all s: this assumption captures the fact that the fox does not have any preconceptions about the position of the grapes and the like. It is supposed that the modeler and the fox agree on the fox's initial probability and utility functions: that is, they represent the fox's attitudes in the first situation, *according to both the fox and the modeler*. The modeler and the fox will disagree however on the representation of the fox's attitudes after his attempt at getting the grapes. For the modeler, the effect of the first attempt can be represented as a change of probability to a new function p_2: thinking of it this way, the fox learns from his first attempt (that it is more difficult than he thought to get the grapes). After the change, the fox's expected utility thus becomes $\sum_{s\in S} p_2(s)u_1(f(s))$. However, as opposed to the case of the previous analysis, it *is* always possible to rewrite the expected utility formula *as if* there was a change in the utility and *not* in the probability. One obtains the representation by $\sum_{s\in S} p_1(s)u_2(s, f(s))$, where $u_2(s, c) = \frac{p_2(s)}{p_1(s)}u_1(c)$. (Note that u_2 is a state-dependent utility: it is a function of states and consequences; see Section 2.1.2.) This is the sort of change that the fox *sees himself* as undergoing: according to him, he has not learnt that the grapes are harder to obtain, he has just changed his mind about whether he wants them or not. As for the analysis in Section 2.2.1, this is only a general schema: different concrete models are obtained by adding particular theories of belief change.

Properties of the analysis

What has changed Concerning the question of what has changed, the point of view taken on the situation is crucial. All are agreed that the expected utility has altered; however, whereas the theorist's representation traces the change to a change in beliefs, the fox represents the change to himself as stemming from an alteration in his

utilities. Under this analysis, sour grapes does not pose a *specific* problem for the modeler: it can be modelled with ordinary belief change apparatus. Sour grapes is merely a phenomenon of self-justification, and, at this stage at least, only a change in the attitudes one considers oneself to have.

The nature of the change There are two aspects of the change (from the modeler's point of view): firstly, the experience of the first attempt at obtaining the grapes causing a change in beliefs, and secondly, a reluctance to recognise the change in expected utility as ensuing from a change in beliefs. Given that neither of these factors are intentional in themselves (the first attempt is intentional, its result, and the belief change caused, is not), the change comes out as causal rather than intentional. Hence interpretations of belief change mechanisms as models of "unintentional" change are pertinent here; mechanisms which support such interpretations apply more naturally.

The source of the change The role of the first attempt, and more particularly the influence of past states and choices on the preferences in the second situation, indicates that there is (diachronic) state-dependence. Moreover, the fact that the same states and consequences are involved, and the same acts are on offer, in the first and second situations implies that this analysis does not consider the changes as possibility-dependent. It thus cannot account for the subtle relationship between state- and possibility-dependence.

The permanence of the change From one point of view, this analysis seems amenable to reversibility: since the utility of the fox "really" remains the same (from the modeler's point of view), it is no surprise if he "reverts back" to this utility. However, according to this analysis, the changes in preference are due to changes in belief; therefore it is the theories of belief change which will have to explain the observed changes in preference, and notably the reversibility phenomenon. The capacity of theories of belief change to account for the reversibility may depend on the theory considered.

According to many major theories of belief change, changes in belief are not in general reversible; thus the short- and long-term reversibility may not be accounted for by versions of this analysis which use such theories.[10] This is the case for the most important quantitative model of belief change, Bayesian conditionalisation: under conditionalisation by an event, information about the probabilities of events which are incompatible with the conditionalising event is lost. Similarly points can be made for logical theories of change whose operators are analogous to conditionnalisation: public announcement logic is an example (van Ditmarsch et al. 2007). Although such theories do not easily account for reversibility, they do naturally capture the long-term change in preferences, in terms of the long-term changes of beliefs brought about by, say, conditionalisation.

On the other hand, reversibility is allowed by the generalisation of Bayesian conditionalisation proposed by Jeffrey (1972): one revises by modifying the

[10] This indeed is in harmony with Elster's use of the reversibility phenomenon to distinguish sour grapes from learning (Elster 1983, pp112–114).

probabilities of events in a partition, without generally setting any of the probabilities to zero or one, so that one can reverse the change by setting the probabilities back to their initial values. Similarly, traditional AGM theories of belief revision (Gärdenfors 1988) exhibit reversability insofar as the operation of expansion, by which new information consistent with current beliefs is incorporated, followed by the operation of contraction, by which beliefs are removed from the agent's corpus, yields the initial set of beliefs.[11]

Points similar to those made in Section 2.2.1 concerning the proper representation of the trigger for change in the formal model hold here: if, immediately following the fox's decision to walk away, there arose an opportunity to get the grapes (he spots a ladder, for example), is this to be thought of as a retraction of some belief he had acquired about the states of the world (a contraction) or the acquiring of the new information that the probabilities of the states of world were as he had originally thought (as in the case of Jeffrey conditionalisation)?

This analysis has the advantage over the previous one that there is a larger amount of work on belief change to draw upon. Furthermore, whereas several belief change operations cannot adequately capture some of the subtle properties of sour grapes, in particular relating to reversibility, others seem more capable. Nevertheless, the analysis does share some of the potential difficulties of the previous one: the relation between state- and possibility-dependance is not accounted for, and subtle interpretation of the belief change operations is required to account for the non-intentional nature of the change. Once again, it is unclear whether these are to be taken as indications that sour grapes is not a case of belief change, or simply as a challenge to be overcome by theories of belief change.

Moreover, there are several other aspects of the model which some might find unsettling. There is a certain intuition according to which there is no change of beliefs involved in sour grapes: the beliefs of the fox in this model concern the position of the grapes, the height of the fox and similar information, and he knew all of this information *before* his first attempt. So what has he learnt? The natural answer seems to be that he has learnt the chances of success at obtaining the grapes, given that they are at such a height. As just noted, these do not correspond to beliefs represented by his probabilities over states of the world. Rather they correspond to a fact about the decision situation he is in: namely, to how reliably the acts on offer effectuate the transitions from states to the consequences which they claim to. This is a guiding intuition for the third analysis.

2.2.3 Reliability of Acts

The analysis Under the final approach to the phenomenon of sour grapes, the fox does not learn anything about the states of nature, nor does he alter his utility for

[11] It seems that the same cannot be said of dynamic logic approaches to belief revision, such as that developed by Baltag and Smets (2006), for they do not have a contraction operation.

consequences, but he *alters the way he represents the decision problem he is faced with*. A concise way of representing this change is by the addition of a *situation-dependent* or *context-dependent* factor in his calculation of expected utility. This factor can be left explicit, where it receives a natural interpretation as a change in his opinion about the reliability of the acts on offer; on the other hand, it can be absorbed into the utility function and thus be interpreted as a change in it.

Formally, the simplest proposal is to introduce a real-valued function on the pairs of states and consequences (s, c), call it γ, which features in the representation of preferences in the second situation. For initial probabilities and utilities p_1 and u_1, the preferences in the first situation are represented by the ordinary expected utility formula, whereas the preferences in the second situation are represented by $\sum_{s \in S} p_1(s).\gamma(s, f(s)).u_1(f(s))$. Just as the other sections presented analysis schemata to be filled in with theories of utility change and belief change respectively, this is but a type of analysis: a concrete analysis is attained by adding a theory of the factor γ.

There are several possible interpretations one could give of this factor. A natural one was mooted above: γ is a *reliability factor*, which reflects the chances of success of an act purporting to take a given state to a particular consequence. If $\gamma(s, c) = 1$, an act purporting to take state s to consequence c will certainly succeed, if $\gamma(s, c) = 0$ it will certainly fail and for intermediate values it will have intermediate chances of success (assuming γ to be normalised to take values in $[0, 1]$). As such, γ can be thought of representing *constraints* on whether acts can deliver particular consequences given particular states: it is thus something which is built into the agent's representation of the decision problem he is facing.[12] This is the principal difference with respect to the analyses of sour grapes proposed above. As remarked in Section 2.1.1, the agent separates the possible consequences of his actions, the properties of the world which determine whether these consequences are reached, and considers the options to be functions taking these possibilities to the appropriate consequences. Under the first analysis, it is the utilities of the consequences which change after the fox's first attempt at getting the grapes. Under the second analysis, it is the beliefs about the world. Under the current analysis, it is the very structure of the decision problem which the agent is assuming that changes. The factor γ is thus properly thought of as a *context-dependent* or *situation-dependent* factor which reflects aspects of the decision problem; the sort of change currently under consideration is a change in the agent's perception or representation of the choice he is facing.

Remark 2 Naturally, the way the agent represents the decision problem encapsulates the things which he presupposes to be true of that problem. To the extent that these *presuppositions* can be qualified as beliefs, the change can be thought of as a sort of change of belief (for a model of presuppositions, and a discussion of its relation to explicit beliefs, see Hill [2008b]). However, as noted above, this change in

[12] Note that it is traditional practice in economics to build the appropriate constraints into the decision problem.

belief is not a change in his beliefs about the states of the world (which in this case specify the position of the grapes) and so this is not a change in his probabilities, as was the change discussed in Section 2.2.2. Of course, the line between explicit beliefs and presuppositions varies with the representation of the problem. It is possible to represent the decision problem so that the issue of the reliability of acts is left open: it suffices to use states of the world which specify the reliability of the acts carried out in them (for example, in a given state of the world, the grapes are at a height of 2 m and the act of grasping them is reliable, in another state, the grapes are at 2 m, and the act is not reliable, and so on). If the decision problems for the fox's first and second attempt are represented like this, the presupposition about the reliability of acts becomes an explicit belief, represented by probabilities, and the change is of the same sort as in Section 2.2.2. However, the sorts of change which are of interest here, and which differ from those discussed in the previous sections, are changes in the representation of the decision problem (or, in other words, changes in the presuppositions). The use of the factor γ is merely one way of representing such a change. It is, however, not the only one.

Another possibility is to represent the change as involving a change in the set of states – and therefore the set of acts – in the representation of the decision problem and the introduction of a probability function on the new set of states which is suitably related to the probability function on the old set of states (in particular, such that they agree on common events). For example, the set of states used up to now in this section will be replaced by states mentioned in the previous paragraph which specify not only the position of the grapes but also the reliabilities of acts. Models of such a change can be supplied by models of states of belief which can cope with awareness and awareness change (for example, Hill [2007b, 2008a], or in a probabilistic setup Modica [2008]).

A third possibility is afforded by the Jeffrey framework. (As elsewhere in the paper, we concentrate attention on Jeffrey's evidential decision theory, without taking sides in the debate which opposes it to causal decision-theoretic variants.) Recall that, under the assumptions that the probabilities of states are act-independent and that the utilities of consequences are act-independent, we have $u(F) = \sum_{S_i} \sum_{C_j} u(C_j)\, p(S_i)\, p(C_j/F \wedge S_i)$ (see Remark 1 for discussion and notation).[13] Recall that one obtains Savage's expected utility formula under the assumption that the third term in the product is equal to one when $f(s_i) = c_j$ and is zero otherwise. Weakening this assumption introduces a third term into the traditional expected utility formula and is thus reminiscient of the introduction of the factor γ proposed above. Hence, in the Jeffrey framework, the change in preferences can be represented by the fact that this third term ceases to take the standard values assumed in the traditional expected utility formula. In other words, the change is a change in the probability of reaching the consequence given the state and the act which purports

[13] It was noted in footnote 5 that in the Jeffrey framework it is natural to remove the states, by collecting the $p(S_i)\, p(C_j/F \wedge S_i)$ terms into a single $p(C_j/F)$ term. Doing this removes the probabilities of states of the world and hence obscures the distinction between the analysis discussed here and that of the previous section. For this reason we do not consider this option further here.

to yield that consequence on that state. Naturally, this is very close to the reliability interpretation of the factor γ. Furthermore, just as for the proposals in the previous paragraphs, under this analysis the change does consist in a change of belief, but not a belief about the states of the world: it is rather a belief about the decision situation which the agent is faced with, and more particularly about the real effects of the options on offer. If the change is understood this way, theories of change in these conditional beliefs, such as that proposed by Bradley (2005), can be used.

Our purpose here is to draw attention to the presuppositions or factors which reflect the decision situation the agent considers himself to be in and to the fact that they change and could be at the root of some preference changes. We do not claim that the above considerations provide anything close to a comprehensive theory of the relevant aspects of the decision problem, nor a comparison of possible representations of them. Indeed, we do not even claim that the three proposals discussed briefly above are equivalent. It is evident that they are not: for example, whereas the first proposal always yields preferences which satisfy Savage's most fundamental axiom on preferences – the sure-thing principle – the last one does not.[14] Indeed, in the light of differences such as these, it may be that the most appropriate representation of the situation-dependent factor will depend on the framework which one is using (see also Remark 3, Section 2.3.2). Given that the framework adopted here is Savage's, and the factor γ seems to be the option which fits in easiest and most simply with this framework, we shall focus on this option for the rest of the paper. The discussion is intended to carry over to other representations of the situation-dependent factor, although it remains to be seen on a case by case basis to what extent it does.

Properties of the analysis

What has changed Note firstly that, like the first analysis (Section 2.2.1), it is not always possible to reformulate the change in the fox's preferences as if it consisted in a change in belief regarding the states of the world.[15] However, it *is* always possible to reformulate it as if it were a change in utilities: $\gamma(s,c).u_1(c)$ may be thought of as a utility function. (Note this utility function is state-dependent.) In this sense, the experience can be thought of as causing a *change in the utility function* from the initial utility u_1 to $\gamma(s,c).u_1(c)$; what is more, γ characterises exactly this change. Evidently, it is *not* necessary to see this as a utility change, because $\gamma(s,c).u_1(c)$

[14] For the former claim, see Section 2.3. As for the latter claim, given the presence of non-zero terms $p(C_j/F \wedge S_i)$ where $f(s_i) \neq c_j$, there will be utility contributions from pairs s_i, c_j where $c_j \neq f(s_i)$, and this contradicts the sure-thing principle. Changes in the set of states may also produce violations of the sure-thing principle with respect to the initial states, if one translates the acts on the initial set of states to acts on the new set of states in the appropriate way; see Savage (1954, §5.5) and Hill (forthcoming, §4).

[15] This will only be possible in the degenerate case where γ is independent of its second value c, and where it is a probability measure on S. See also Remark 2.

is not the only utility involved in second situation: there is still the initial utility u_1. This is, so to speak, the *absolute* utility, independent of the situation, whereas $\gamma(s,c).u_1(c)$ is the utility *in this situation* – relative to the situation insofar as the situation limits, through the factor γ, the accessibility of the consequence c, or the chances of actually obtaining this consequence.[16]

Several of the attractive properties of the analysis come from the distinction between the utility "effective" in the decision and the agent's "underlying" utility. The first of these properties has already been evoked: in this analysis, sour grapes comes out as a *change in the situation-relative utility*, though not in the *absolute utility*.[17] This has interesting consequences for utilitarianism: should one use absolute utilities in considerations of social good or more variable situation-relative utilities? The argument against utilitarianism mentioned at the beginning of the paper is evidently stronger in the latter case. Moreover, the distinction between the utility which the agent effectively employs and some more stable underlying utility is reminiscent of other distinctions which have been recently proposed in decision theory, such as the distinction between "experienced" and "intrinsic" utility argued for by Kahneman et al. (1997). Naturally, one might either interpret the fox as not being conscious of the duality of utilities, and thus as thinking that his utility has changed, or as being lucid as to the pair of utilities, and thus aware, when he mumbles that the grapes are no longer desired, that his affirmation is context-relative and refers to the situation-relative utility, not the absolute utility.

The nature of the change As noted above, the change is a change in the representation of the decision problem which the agent is faced with. Some of the models for changes of this sort only admit natural interpretations according to which the change is non-intentional. Above it was noted that the change could be captured by a change in the set of states: this would involve an increase in the awareness of the agent, and one cannot intentionally decide to become aware of something (Hill 2007b). Similarly, there is a sense in which it sounds strange to intentionally question a presupposition (one is usually "forced" or "led" to question it). On the other hand, some models support intentional as well as non-intentional readings, and the latter are required here. For example, it was noted that, in the appropriate framework, the change can be thought of as a change in certain types of conditional beliefs; as for the literature on belief and utility change (Sections 2.2.1 and 2.2.2),

[16] To an extent, this duality in utilities can be reproduced under the other two proposals mentioned in Remark 2. In the Jeffrey framework, the situation-relative utility is $u(C_j \wedge F)p(C_j/F \wedge S_i)$; note that, unlike the proposal in the Savage framework, this utility is not only state-dependent, but also act-dependent. As for the change in the set of states, such a change is equivalent to a change in the set of consequences (see the "small world consequences" in Savage [1954, §5.5]); the utility for these new consequences can be thought of as the situation-relative utility.

[17] The interpretation of the difference of these utilities is not unrelated to the interpretation of the factor γ. The fact that, in general, γ is a way of capturing an aspect of the decision situation the agent sees himself as faced with justifies the terminology 'situation-relative utility'. In the particular case that γ is considered to reflect the reliability of acts, one could give this utility another name, such as 'reliability-discounted utility'. For other interpretations of the difference between the two sorts of utility, see Hill (forthcoming).

most of the discussions of such changes consider examples where the incoming information is in the form of new conditional beliefs which are to be intentionally incorporated into one's corpus (Bradley 2005).

The source of the change In understanding the factor γ as encapsulating an element presumed to be integral to the decision problem the agent perceives, one can easily take account of the flexible distinction, noted in Section 2.1.2, between the state- and possibility-dependence of the change. On the one hand, the change between the two situations can be considered as a result of the choice in the first attempt and his experience of the results of that choice: there is thus diachronic state dependence. On the other hand, it was noted above that the factor γ represents constraints on the range of acts he can expect to carry out successfully. In other words, the introduction of this factor implies a change in the possibilities *effectively available* to the fox in second situation. In this sense, there is possibility-dependence.

The permanence of change The fact that the belief and absolute utility are left fixed, but that some sort of situation-dependent factor intervenes, allows a rather intuitive understanding of short-term reversibility. The choice of walking away from the grapes, and the choice of using the ladder which the fox notices as he walks away to get them are made in different situations, and it certainly seems that the chances of success of the act of attempting to get the grapes differ between these situations. These are two different decision problems, involving different situation-dependent factors. This is but an intuition: naturally, an adequate theory of the representation of the decision problem or of the situation-dependent factor, which accounts for such differences in an appropriate way, is required.

As for the long-term permanence of the preference change, this may perhaps be modelled by the fact that, throughout an extended period, there is a generally similar sort of decision problem, with a similar set of acts and a stable situation-dependent factor; once again, such a model remains to be developed formally. Such a model, combined with the duality of the utilities, would allow an understanding of the intuition that it is the agent's utility that changes over such extended periods, by reasoning as follows. Since the situation-dependent factor is generally stable over the period, the same situation-relative utility applies throughout the period. If one comes to presuppose on this basis that the situation-dependent factor is fixed at that value, the situation-relative utility, which integrates this factor, does not vary during the period and can be effectively treated as the "real" utility in decision problems during this period, usurping the absolute utility. It would thus seem that the intuition that a change of utilities is involved in such cases actually refers to the difference between the absolute utility (which is the same before and after the change in preferences) and the situation-relative utility (which appears to be constant because of the stability of the sort of decision problem faced). Given this account of the long-term permanence of the preference change, long-term reversal of preferences could be understood in terms of a change of circumstance which is drastic enough to invalidate the presupposition that the situation-dependent factor is fixed – that is, by a decision problem with a different situation-dependent factor – in such a way that

resort to the pure utility is once again required.[18] Under this account, long-term reversibility is a similar sort of effect to short-term reversibility, though perhaps of differing degree.

Just as for the other analyses, this is more a sketch than a fully developed analysis, insofar as a fully worked-out account of the representation, appearance and dynamics of the situation-dependent factor is still lacking. Nevertheless, depending on one's model of this factor, this analysis sits more or less easily with theories of change which deal explicitly with cases of non-intentional change, it is able to account for the flexible distinction between state- and possibility-dependent factors and it does seem to be subtle enough to cope with the reversibility and permanence phenomena.

2.3 Getting Your Teeth into Sour Grapes

What changes in a case of sour grapes – utilities, beliefs, or some aspect of the decision problem as the agent perceives it? An indirect method for answering this question would identify properties of sour grapes which, according to theories of attitude change, only occur in, say, changes in utility. It is fair to say, on the basis of the considerations in the previous section, that this method does not yield any strong conclusion: many of the doubts which could be emitted with respect to the analyses, such as the question of the intentional nature of the changes and to a certain extent the reversibility phenomena, seem to be as much challenges for theories of belief or utility change as they are specific features of one or other sort of change. The apparently more substantive differences between the analyses, such as the ability of only the third analysis to capture the subtle relationship between state- and possibility-dependence, seem too slight to support a firm endorsement of one at the expense of the others. To answer the essential question of sour grapes, a more direct method is required.

Such a method cannot rely solely on theories of attitude change, because these theories generally assume that one knows the attitude which is to change, and go on to say how it should change; that is, they assume that the answer to the central question of sour grapes is already known. Something else is needed: a way of determining what the agent's attitudes are at any given moment. Given this, one could see what has changed by eliciting the agent's attitudes before and after the change and comparing them. Note that this method applies most naturally to *instances* of preference change, rather than to the *class* of changes which can be classified as cases of sour grapes: indeed, given that the properties of the sour grapes phenomenon are not sufficient to deduce what sort of change are occurring, it is possible that in different cases of sour grapes-style phenomena, different sorts of changes are taking place. Throughout this section, an arbitrary particular instance of preference change will be considered.

[18] Indeed, in Elster's example of long-term reversibility (of the preference for city life; see Section 2.1), a drastic change is required (a move to the countryside for a considerable period).

To the problem of change which has been considered up to now – the problem of understanding the relationship between the agent's attitudes *in different situations* – we thus add the problem of *elicitation* – that of distinguishing the part of the agent's preferences which is due to his beliefs from the part which is due to his utilities *in a given situation or decision problem*. To answer the question posed by sour grapes, one requires a reply to both problems at once: a way of distinguishing between the agent's beliefs and his utilities in a given situation *which is such that* it yields an understanding of the changes they undergo.

2.3.1 The Direct Method: The Classic Approach

Decision-theoretic preliminaries The question of elicitation has not been a subject for logicians, but for decision theorists. At the heart of every decision theory is a *representation theorem*, giving a set of necessary and sufficient conditions on the agent's preferences for there to exist a unique probability and an essentially unique utility representing the preferences (according to the expected utility formula, in the present case). The uniqueness of the probabilities and utilities is crucial: it allows proponents to claim that they represent the beliefs and desires underlying the agent's preferences. Related to these theorems, or their proofs, there are often methods for practically eliciting or approximating the agent's probability and utility functions from their choices. The conditions in the representation theorems become assumptions that are needed for the success of the elicitation techniques. (Beyond the practice of behavioural decision theorists, this sort of connection between norms on rational preference and the conditions for the possibility of understanding the agent by attributing attitudes to him is present in philosophy, notably in the work of Donald Davidson.) At least this is so with the most developed and simplest decision theory, namely the theory of expected utility pioneered by Savage, and we shall focus almost entirely on this theory for the rest of the paper.[19]

Representation theorems such as Savage's generally work with a single decision situation: a single set of states, consequences and acts, with a single preference relation over the acts. Although the axioms in the theorem are formulated as *conditions on the preference relation*, they correspond more or less neatly to *principles concerning the rationalisation of the agent's preferences*. In endorsing the conditions on preferences as norms on rational preferences, one is committed to endorsing the principles governing the rationalisation of his preferences, and vice versa. Given that the details of the axioms on preferences are not important for the purposes of this paper, we present the underlying principles involved in Savage's theorem, with only very rough indications as to the corresponding condition on preferences; for precise formulation and discussion, see Savage (1954).

[19] For some comments concerning the application to the Jeffrey framework, see Remark 3 below.

Order [P1] The preference relation is representatable by a real-valued function on the set of acts. (Condition on preferences: the relation 'is preferred to' is a transitive and complete order.)

Independence [P2, also called the Sure-Thing Principle] The aforementioned real-valued function on acts can be decomposed into the sum $\sum_{s \in S, c \in C} V(s, c)$ where V is a real-valued function on state-consequence pairs. (Condition on preferences: the preference relation over two acts depends only on the states where they differ.)

State-independence [P3 and P4] The function V can be decomposed into a probability p on states and a state-independent utility u on consequences. (Condition on preferences: for any pair of consequences, the preference order over them given any event is the same; for two-valued acts, the preference relation over them depends only on the set of states where they take the more preferred consequence, and not on the consequences themselves.)

Continuity [P6 and P7] The functions mentioned at the previous levels exist for large sets of states and are suitably unique. (Condition on preferences: the set of states and the preference relation are sufficiently rich and well-behaved under taking limits.)

For any agent satisfying these principles, and hence for any agent whose preferences satisfy the corresponding axioms, there exists a unique probability over states p and a suitably unique[20] state-independent utility u such that he prefers an act f to g if and only if the expected utility of the former is greater than that of the latter: $\sum_{s \in S} p(s).u(f(s)) \geq \sum_{s \in S} p(s).u(g(s))$.

A detailed discussion of these principles and the corresponding axioms on preferences is beyond the scope of the paper (see for example Savage (1954); Broome (1991)). Although there is reason to question each one, it is safe to say that the first two are often taken to have some normative justification and the last is largely considered to be "technical" or "structural". As for the second last (state-independence), it has been challenged by decision theorists such as Karni and Drèze, who have proposed alternative principles. It is worth underlining that none of these principles can be dropped without being replaced: each is necessary for the representation theorem and the possibility of eliciting attitudes. For example, the theorists mentioned above who challenge state-independence but retain the Savage framework have had to propose alternative, more complicated principles such that it is possible to elicit unique probabilities and state-dependent utilities from any agent who satisfies the new principles.

The common strategy and its weaknesses The natural strategy for determining what has changed in a particular instance of sour grapes employs theories of decision such as the one described above in the following way: assume that the agent's preferences satisfies the conditions for elicitation before the preference change, and use the elicitation methods mentioned above to elicit his probabilities and utilities;

[20] Precisely: unique up to positive affine transformation.

assume the agent's preferences satisfy the conditions after the preference change, and elicit his (new) probabilities and utilities; finally, compare the two sets of attitudes to see what has changed. We shall argue that traditional elicitation techniques which are more or less loosely related to representation theorems such as Savage's are not propiscious for use in such a strategy. The norms and conditions which guide the elicitation are entirely synchronic, and hence surrender any pretension of remaining faithful to features of the dynamics of the agent's attitudes.

The point can be made on the case of sour grapes. Suppose that the fox satisfies the principles above, and hence the corresponding preference axioms, in the first situation (when he decides to try to get the grapes) and in the second situation (when he walks away); hence the typical representation theorems of decision theory – Savage's theorem, for example – apply and his probabilities and utilities can be elicited. Suppose furthermore that this method of elicitation yields probability and utility functions according to which the first analysis (Section 2.2.1) is correct: the probabilities elicited in the two situations according to Savage's theorem are the same but the utilities have changed. Suppose finally that as the fox walks away from the grapes, he spots a ladder, seizes it, clambers up and grabs the grapes. This situation seems naturally understandable in terms of his high utility for the grapes (relative to not having them). Indeed, assuming that the agent satisfies the above principles (and hence the corresponding preference axioms) in this third situation, his attitudes can be elicited, and it indeed turns out that he has the same high utility for grapes which he had initially. Now we have a clash of intuitions. On the one hand, this sequence of situations and actions constitutes the (short-term) reversibility phenomenon recognised in Section 2.1.2; as noted in Section 2.2.1, it is unnatural to explain this phenomenon by a pair of sudden changes in utility. On the other hand, the strategy under consideration for deciding what has changed in an instance of sour grapes, which uses the repeated application of Savage's representation theorem and related elicitation techniques, implies that there is just this erratic pair of utility changes. What is going on?

Generalising from this example, consider two decision situations σ_1 and σ_2, perhaps with different acts on offer but with either a common set of possible consequences of these acts (so the states may be different) or a common set of relevant states of the world (so that the consequences may be different). The tension occurs when (1) there is a strong intuition that the agent's utilities (respectively beliefs) are the same in σ_1 and σ_2 but (2) he satisfies the conditions required in Savage's representation theorem in both σ_1 and σ_2, and the elicited utilities (resp. beliefs) differ. The previous paragraph contains an example of this sort for utilities (with σ_1 being the second situation in the story, and σ_2 the third); for a numerical example involving beliefs, see the interpretation Hill (forthcoming) offers of an example proposed by Karni (1996, pp 256–257). In examples such as these, one could ignore the intuition about the stability of attitudes, and follow the results of the elicitation blindly. However, the tension seems to indicate that there is an element of rational behaviour which is not captured by the principles underlying results such as Savage's. Namely, a rational agent's attitudes should not change gratuitously between appropriately related decision situations. Let us call this the *stability principle*.

Stability In the absence of a reason for the agent's probability or utility to differ between suitably related decision situations, they remain the same.

Naturally, like our rendition of the principles underlying Savage's theorem, this principle is formulated in terms of the rationalisation of the agent's preferences, rather than as a direct condition on the preference relation. As discussed briefly below, this principle can be translated into precise (behavioural) conditions on preferences, which feature in a representation theorem and serve as necessary and sufficient conditions for the success of an elicitation procedure. Under the translation, the concepts featuring in the principle as stated above are sharpened. For example, the conditions on preferences imply that the set of suitably related decision situations between which there is no reason for the agent's probability or utility to differ has certain properties; as such, they will provide a minimal axiomatic characterisation of this notion. Of course, in practice, it is up to the good judgement of the elicitor to decide what counts as an appropriate related situation, just as it is up to him to provide an adequate representation of the decision problem (the sets of states and consequences, in the case of Savage). Further discussion of the notion of suitably related decision situation is beyond the scope of this paper; the reader is referred to the discussion in Hill (forthcoming).

The examples considered above show that stability and the four aforementioned principles of classic Savagean decision theory may enter into conflict. In hindsight, this is not at all surprising. The Savagean axioms, like all axioms in decision theory, are synchronic: they deal, at least initially, with attitudes and choices in a given, fixed decision situation. By contrast, the stability principle is diachronic: it explicitly takes into account the relationship between decision situations. One cannot expect synchronic principles to be able to account for diachronic properties; indeed, the methods built on the Savagean principles fail to yield an intuitive understanding of the diachronic behaviour of the agent's beliefs and utilities in some cases. Do we have reasons for accepting the diachronic stability principle? And what is the price of accepting it?

Why stability? The first argument in favour of the stability principle comes from our intuitions. How do you tell whether the fox has *really* changed his mind about the desirability of the grapes? See how he acts in other situations where they are more accessible (if they were offered on a plate for example, or if, just after turning away, he spots a ladder). Such folk wisdom invokes the stability principle in the two senses mentioned above. Firstly, as a norm: a rational agent should not change his utilities between such appropriately related situations. Secondly, as a guide to the agent's attitudes, and thus a way of understanding his behaviour: to elicit his utilities as he walks away, it is sufficient to elicit them in the related situation where, at just that moment, he spots a ladder. We do seem to use the stability principle in our rationalisations of human behaviour.

A second consideration in favour of the stability principle relates to the general project of understanding change. To talk of change, one must be able to make sense of what it means for there to be lack of change. This is generally the "null state" on

which theories of change build. Without the stability principle, attitudes risk being too erratic to allow such a null state. This was seen above in examples where, using elicitation methods which ignore inter-situational comparisons, utilities differ between situations where there is supposed to be stability. With such chaotic behaviour of attitudes between situations, meaningful investigation of attitude change becomes well-nigh impossible. If the problem of change is to be treated as well as that of elicitation, the principles underlying the understanding of an agent's behaviour and the elicitation of his attitudes must accommodate the basic requirements for an investigation into change. The stability principle is a way of doing this.

Regarding the price to pay for acceptance of this principle, note that the stability principle does not contradict the four standard Savagean principles. It may well be that the agent satisfies the Savagean principles in all decision situations, and that the probabilities and utilities elicited using standard techniques are the same in suitably related decision situations. However, the examples above seem to indicate that there may be cases where the Savagean principles and stability cannot be simultaneously respected. If the stability principle is to be retained as a norm for rational choice, then one of the Savagean principles will need to be sacrificed. Which one?

Stability only involves the probabilities and utilities of the agent. Hence it does not interfere with the order condition or the independence condition, which feature in the representation theorem, and implicitly in the process of elicitation at a stage before probabilities and utilities have been separated out. The two most important principles of Savagean decision theory are thus fully retained on adoption of the stability principle. Putting aside continuity, which is technical and does not explicitly involve the probabilities and utilities, all that remains is state-independence; and indeed, this is the condition which traditionally allows one to separate the probability from the utility. If stability is adopted as a guiding condition for elicitation of attitudes, then it is in place of state-independence. However, the normative and descriptive validity of the axioms for state-independence, as well as the principle stating that the agent's utilities are always state-independent, have been doubted, both by economists (Drèze 1987; Karni and Mongin 2000; Arrow 1974) and philosophers (Joyce 1999), so much so that it is fair to say that there is a consensus that, although they hold in some decision situations, they surely do not hold in others. Indeed, for many of these authors, they do not in general constitute an acceptable norm for rational decision, and thus cannot be assumed in the elicitation of attitudes. Above, when arguing that traditional methods are inadequate for the elicitation of attitudes in cases of change, it was assumed that all the Savage axioms held in all situations, so that the traditional techniques always work. However, the situation is in fact worse for the defender of traditional representation theorems and elicitation methods: in some situations, the conditions on preferences do not hold and the techniques cannot even be applied. The stability principle replaces the weakest plank in the Savage construction.

2.3.2 An Alternative Approach

Using stability Stability says that the agent's utilities (respectively probabilities) are the same in appropriately related decision situations. This allows a simplification in one's elicitation of the agent's attitudes. To elicit his utilities in the current decision situation, one can elicit his utilities in any appropriately related situation: assuming stability, the result will be the same. So choose the situations which are the easiest for elicitation: not only will the use of these situations make the elicitation easier, but by consequence it will be rendered more reliable.

Consider once again the sour grapes example, and consider the task of eliciting his utilities as he walks away from the tree. This is a complicated case for the elicitor: the fact that he does not choose to attempt to get the grapes is not an indication that he does not value them, because his beliefs may play an important role. It is much easier to elicit his attitudes in a situation where the grapes are offered on a plate or where a ladder is available: in these cases, if he chooses to get the grapes, this is an indication that he values them, because the beliefs about the relevant issues in the choice are also easy to elicit. Indeed, even if there were a sophisticated procedure for eliciting attitudes using the preferences in the former situation alone, it is more prone to error and more likely to yield counter-intuitive results than the use of preferences in the latter situation. The recourse to other appropriate situations, permitted by the stability principle, is a more robust way of eliciting preferences.

These considerations yield an elicitation method in which the stability principle plays a central role. We will say that a decision situation is *simple* if the classic elicitation methods apply uncontroversially in this situation; in particular, one expects the decision-maker to maximise expected utility and his utilities to be independent of the states. To elicit the probability and utility of the agent in any given decision situation σ, find another pair of decision situations σ_1 and σ_2 such that (1) σ_1 has the same set of states as σ and σ_2 has the same set of consequences as σ; (2) σ_1 and σ_2 are simple; (3) σ_1 and σ_2 are suitably related to σ in the sense of the stability principle, so that the principle can be taken to apply. To elicit the agent's probabilities in σ, elicit his probabilities in σ_1 using traditional methods (these are robust since σ_1 is simple). By stability, the probabilities elicited in σ_1 are also the agent's probabilities in σ. The agent's utilities in σ can be elicited in a similar way, by eliciting the utilities in σ_2.

This elicitation method, based as it is on stability, does not suffer from the objection to the previous one. If the fox, as he walks away from the grapes, spots the ladder, grabs it and clambers up to get the grapes, this choice is a factor in the elicitation of the attitudes he had as he was walking away. Since there is no reason for him to change his utilities when he spots the ladder, by stability, it can be assumed that they are the same when he walks away (the second situation in the story) and when he returns with the ladder (the third situation). Since the latter situation shares the same set of consequences (obtaining the grapes or not) and is simple (the utilities of the grapes are easily read off from his choices), the utilities elicited in this situation provide a reliable indication of his utilities as he was walking away. According to this process of elicitation, his utilities *are* the same in throughout the story, as the

intuition concerning short-term reversibility would suggest. These considerations, taken in tandem with the arguments above in favour of the stability principle, suggest that this is a more adequate method of elicitation in cases where change of attitudes is of interest.

As noted above, a theory of decision should vaunt a representation theorem, which, in many cases, is the theoretical version of the intuitions at work in the elicitation method. Here is no exception: a representation theorem based on exactly this elicitation technique is presented and defended in Hill (forthcoming). Detailed presentation of this result goes beyond the scope of the paper; let us nevertheless draw attention to several aspects which are relevant here. First of all, as anticipated above, the main principles of Savage's theorem are retained: it is assumed that order, independence and continuity hold in all decision situations. By contrast, state-independence is not assumed to hold in all situations; indeed, in the theorem, the simple situations are formally defined as those where the state-independence axioms do hold (and hence, where traditional results apply). The result thus goes beyond Savage in that it can deal with cases where state-independence fails to hold, and it does so by referring to situations where it does hold. As noted above, there are independent arguments against the general validity of state-independence.

The axioms which replace it are "diachronic", and concern the set of decision situations which are suitably related to a given decision situation, in the sense of the stability principle. The first is a richness axiom, requiring that, for any given decision situation, there exists the sorts of situations required to elicit probabilities (resp. utilities), i.e. suitably related simple situations having the same set of states (resp. consequences). The second is a consistency axiom, requiring that if there are several suitably related simple situations which could be used to elicit the probabilities (resp. utilities) in a given decision situation, they give the same result. The consistency axiom is a direct consequence of the stability principle (and the properties of the relation "begin suitably related"): by stability, if simple situations σ_1 and σ_2 are suitably related, the agent has the same probabilities in them; since the situations are simple, these probabilities can be elicited using traditional methods; hence, the probabilities elicited using traditional methods in these two situations must be the same. To this extent, this axiom can be seen as translating into behavioural terms some of the content of the stability principle. The richness axiom can be thought of as a weakening of state-independence: whereas it is largely accepted that the state-independence axioms do not hold in every decision situation, it is equally evident that there are situations in which they do hold, and it is a version of this weaker fact which is required by the richness axiom. For further discussion of these axioms, and their precise formulation in terms of preference orders, see Hill (forthcoming).

Consequences: the situation-dependent factor The stability principle underlies a technique for eliciting the agent's probabilities and utilities which, it has been argued, yields more adequate results than traditional techniques, especially when change of attitudes is at issue. One final question remains to be addressed: what sort of representation of preferences does this elicitation technique provide?

It should be evident that it is not *necessarily* a representation by the traditional expected utility formula (in the Savage framework): $\sum_{s \in S} p(s).u(f(s))$. Consider a non-simple situation σ and appropriately related simple situations σ_1 and σ_2 having the same states and consequences, respectively, as σ. The elicitation method sketched above yields probabilities p and u obtained by applying traditional methods in the latter two situations. There is however no guarantee that p and u represent the preferences in the σ: it could be for example that f is preferred to g in σ, but that $\sum_{s \in S} p(s).u(f(s)) < \sum_{s \in S} p(s).u(g(s))$. For example, if the fox takes the opportunity to grab the grapes when he spots the ladder, one can infer that his utility for the grapes were as high when he was walking away as it was in the first situation when he attempted to get the grapes. If, furthermore, he were offered a bet on the position of the grapes and accepts the same bet before and after his failed attempt, it can be inferred that his beliefs about the states of the world (position of the grapes) have not changed. This would indicate that the utility and the probability in the first situation (where he attempts to get the grapes) and in the second situation (where he walks away) are the same. However, the preferences in these two situations differ, and so cannot both be represented by the expected utility formula involving the elicited probabilities and utilities.

The most natural appropriate form of representation has already been mentioned in this paper: it is the form proposed in Section 2.2.3 which involved the situation- or context-dependent factor γ.[21] Since the Savage principles of order, independence and continuity are satisfied for any situation σ, for any such situation there is a function on state-consequence pairs, call it V, which represents the preferences in σ (Section 2.3.1). As noted above, V may not be equal to the product of the probability and utility elicited by recourse to appropriate simple situations; however, it is always possible, by taking the quotient, to define a factor γ such that $V(s, c) = p(s).u(c).\gamma(s, c)$ for all states s and consequences c.[22] But this just yields the representation proposed in Section 2.2.3: f is preferred to g in σ if and only if $\sum_{s \in S} p(s)u(f(s))\gamma(s, f(s)) > \sum_{s \in S} p(s)u(g(s))\gamma(s, g(s))$.[23] The situation-dependent factor, proposed above as a possible analysis of the phenomenon of sour grapes, comes out as a natural consequence of the elicitation technique.

We thus have a more general motive for introducing and taking account of something of the order of a situation-dependent factor. If one accepts the stability principle as a norm for rational behaviour and a guiding principle for deciding what an agent's beliefs and utilities are, the situation-dependent factor is needed to "fill the gap" between the beliefs and utilities elicited and the agent's preferences in a given

[21] Several other ways of representing changes in situation-dependent aspects were discussed in Remark 2, Section 2.2.3. Discussion of the relationship to the change in the set of states goes beyond the current paper; the Jeffrey framework is discussed in Remark 3 below.

[22] For details, see Hill (forthcoming). Naturally, a rather innocent axiom regarding null events is required to ensure that the quotient is well-defined when it should be.

[23] This multiplicative representation is the most natural, given that V is normally decomposed into the product of probabilities and utilities. Other possible representations can be imagined (e.g. δ such that $V(s, c) = p(s).u(c) + \delta(s, c)$), but it is unclear how they would be understood.

situation. In hindsight, this should not be surprising. If the beliefs about the relevant states and the utilities for consequences have not changed, then the only other option is some aspect of the way the agent represents the decision problem to himself. Similarly, if you elicit the agent's probabilities over states and utilities for consequences using a method which does not rely uniquely on the situation under consideration, and if his preferences are not represented by the expected utility calculated with these probabilities and utilities, then you have discovered that there is an aspect of the situation as the agent represents it which still needs to be taken into account. The situation-dependent factor is just a simple way of representing this factor.

Remark 3 To those with sympathies for Jeffrey's decision theory, the conclusions above may be particularly welcome. As we saw in Section 2.2.3, there are analogies between the situation-dependent factor γ and the probability of the consequence conditional on the act and the state. To the extent that conditional probabilities are, for many, a central difference between Jeffrey's decision theory and Savage's (but see Levi 2000; Spohn 1977), the above argument could be considered as a vindication of the former over the latter.

Some important differences between the Jeffrey and Savage frameworks should however be noted. The main problem of this section – the problem of elicitation – receives a better treatment in the latter than in the former, in at least two respects. First of all, the classic representation theorem in the Jeffrey framework (Bolker 1967) does not yield unique probabilities, nor utilities which are as unique as Savage's. This has been corrected, at the expense of extra structure, in the work of Joyce (1999) and Bradley (2007b). Secondly, the representation theorem and framework do not easily lend themselves to practical elicitation of attitudes: not only have there been no attempts to elicit attitudes using the Jeffrey framework, but in work on the subject, there is little indication as to how this could be done in general (for example, in Jeffrey [1972], there are only indications as to how to elicit probabilities on the set of indifferent propositions, but not on propositions in general [Chapter 7]).

Indeed, insofar as the elicitation method proposed above purports to be a method for measuring utilities (of consequences), probabilities (of states) and the situation-dependent factor, it could be seen as a contribution to alleviating this weakness of the Jeffrey framework. Using simple situations to elicit probabilities and utilities does not depend on the use of the Savage framework rather than the Jeffrey framework. The only element of the representation theorem described above which is specific to the Savage framework is the use of Savage's independence axiom (the so-called sure-thing principle) to get a representation by a function V of state-consequence pairs. However, Bolker's axioms provide a function of this sort for the Jeffrey framework (indeed, they provide more than this). It is thus natural to conjecture that a representation theorem of the sort described above, which elicits unique probabilities, unique utilities and probabilities of consequences given acts and states rather than a γ-factor, can be proved in the Jeffrey framework. Further discussion of the interpretation of such a theorem and of these conditional probabilities is beyond the scope of this paper; we will content ourselves with signalling this as an area for future research.

In conclusion, three things can be at the root of a given change in preferences: a change in utilities, a change in beliefs with respect to the states, or a change in some factor in the decision problem. A particularity of the phenomenon of sour grapes is that it is not clear which of these changes is involved. The properties of sour grapes-style changes have not proved sufficient to clarify this issue. To determine which change is involved in a particular case of preference change, one will have to elicit the agent's beliefs and utilities. And in considering what is the best way of doing so, given that the elicitation is to be used to understand change, the situation-dependent factor turns out once again to play a crucial role. So it cannot be dismissed as an easy solution to the problem of preference change, or as a special trick which is only relevant for the case of sour grapes. It poses a challenge to those seeking to understand preference change: to correctly model the agent's representation of the decision problem, its relationship with other attitudes and the changes it undergoes. It is unclear how a model of decision and attitude change which cannot account for this would ever be complete.

Acknowledgements The author would like to thank Philippe Mongin and Francesca Poggiolesi for inspiring discussion and helpful suggestions, and three anonymous referees for their detailed comments.

References

Arrow, K. J. (1974). Optimal insurance and generalized deductibles. *Scandinavian Actuarial Journal*, 1:1–42.

Baltag, A. and Smets, S. (2006). Dynamic belief revision over multi-agent plausibility models. In Bonanno, G., van der Hoek, W., and Wooldridge, M., editors, *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory (LOFT06)*, pp. 11–24. Amsterdam University Press, Amsterdam.

Bolker, E. (1967). A simultaneous axiomatisation of utility and subjective probability. *Philosophy of Science*, 34:333–340.

Bradley, R. (2005). Radical probabilism and mental kinematics. *Philosophy of Science*, 72:342–364.

Bradley, R. (2007a). The kinematics of belief and desire. *Synthese*, 56:513–535.

Bradley, R. (2007b). A unified bayesian decision theory. *Theory and Decision*, 63:233–263.

Broome, J. (1991). *Weighing Goods*. Basil Blackwell, Oxford.

Cyert, R. M. and DeGroot, M. H. (1975). Adaptive utility. In Day, R. H. and Groves, T., editors, *Adaptive Economic Models*, pp. 223–246. Academic, New York.

de Jongh, D. and Liu, F. (2006). Optimality, belief and preference. In Artimov, S. and Parikh, R., editors, *Proceedings of the Workshop on Rationality and Knowledge*, *ESSLLI* 2006.

Drèze, J. H. (1987). *Essays on Economic Decisions Under Uncertainty*. Cambridge University Press, Cambridge.

Elster, J. (1983). *Sour Grapes. Studies in the Subversion of Rationality*. Cambridge University Press, Cambridge.

Gärdenfors, P. (1988). *Knowledge in Flux : Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, MA.

Hansson, S. O. (1995). Changes in preference. *Theory and Decision*, 38:1–28.

Hill, B. (2007b). The logic of awareness change. In Arrazola, X. and Larrazabal, J. M., editors, *Proceedings of the First ILCLI International Worskop on Logic and Philosophy of Knowledge, Communication and Action*. University of the Basque Country Press.

Hill, B. (2008a). Towards a "sophisticated" model of belief dynamics. Part I: The general framework. *Studia Logica*, 89(1):81–109.

Hill, B. (2008b). Towards a "sophisticated" model of belief dynamics. Part II: belief revision. *Studia Logica*, 89(3):291–323.

Hill, B. (forthcoming). Living without state-independence of utilities. *Theory and Decision*.

Jeffrey, R. C. (1972). *The Logic of Decision*, 2nd edn. University of Chicago Press, Chicago.

Joyce, J. M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge.

Kahneman, D., Wakker, P. P., and Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, 112:375–405.

Karni, E. (1996). Probabilities and beliefs. *Journal of Risk and Uncertainty*, 13:249–262.

Karni, E. and Mongin, P. (2000). On the determination of subjective probability by choices. *Management Science*, 46:233–248.

Levi, I. (2000). Review of *the foundations of causal decision theory*. *The Journal of Philosophy*, 97:387–402.

Modica, S. (2008). Unawareness, priors and posteriors. *Decisions in Economics and Finance*, 31.

Savage, L. J. (1954). *The Foundations of Statistics*, 2nd edn. 1971. Dover, New York.

Sen, A. K. and Williams, B., editors (1982). *Utilitarianism and Beyond*. Cambridge University Press, Cambridge.

Spohn, W. (1977). Where luce and krantz do really generalize savage's decision model. *Erkenntnis*, 11:113–134.

van Benthem, J. and Liu, F. (2007). Dynamic logic of preference upgrade. *Journal of Applied Non-classical Logic*, 17:157–182.

van Ditmarsch, H., van der Hoek, W., and Kooi, B. (2007). *Dynamic Epistemic Logic*. Springer, Dordrecht.

Chapter 3
For Better or for Worse: Dynamic Logics of Preference

Johan van Benthem

Abstract In the last few years, preference logic and in particular, the dynamic logic of preference change, has suddenly become a live topic in my Amsterdam and Stanford environments. At the request of the editors, this article explains how this interest came about, and what is happening. I mainly present a story around some recent dissertations and papers, which are found in the references. There is no pretence at complete coverage of preference logic (for that, see Hansson 2001) or even of preference change (Hansson 1995).

3.1 Logical Dynamics of Agency

3.1.1 Agency, Information, and Preference

Human agents acquire and transform information in different ways: they observe, or infer by themselves, and often also, they ask someone else. Traditional philosophical logics describe part of this behavior, the 'static' properties produced by such actions: in particular, agents' knowledge and belief at some given moment. But rational human activity is goal-driven, and hence we also need to describe agents' evaluation of different states of the world, or of outcomes of their actions. Here is where preference logic has come to describe what agents prefer, while current dynamic logics describe effects of their physical actions. In the limit, all these things have to come together in understanding even such a simple scenario as a *game*, where we need to look at what players want, what they can observe and guess, and which moves and long-term strategies are available to them in order to achieve their goals.

J. van Benthem
University of Amsterdam and Stanford University
e-mail: http://staff.science.uva.nl/∼johan/

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

3.1.2 Logical Dynamics of Information and Belief

There are two dual aspects to this situation. The static description of what agents know, believe, or prefer at any given moment has been performed by standard systems of philosophical logic since the 1950s – of course, with continued debate surrounding the merits of particular proposals. But there is also the dynamics of actions and events that produce information and generate attitudes for agents – and gradually, these, too, have been made a subject of logical investigation in the program of 'logical dynamics' (van Benthem 1996, 2008). For instance, an observation or an answer to a question are informative events that can be put explicitly inside complete systems of dynamic logic, which describe what agents know before and after such events take place. For purposes of exposition, this paper will use the current methodology of 'dynamic epistemic logic' (cf. van Ditmarsch et al. 2007; Baltag et al. 2008), and some concrete systems will be found below. A typical formula of such a system might say the following:

$$[!\varphi]K_i\psi \quad \text{after receiving the 'hard information' that } \varphi, \text{ agent } i \text{ knows that } \psi. \tag{3.1}$$

This describes knowledge of individual agents after direct *information update*, and the account can also deal with complex group scenarios where agents have different observational access to the actual event taking place (think of drawing a card in a game). By now, there are also dynamic logics that describe more subtle 'policy-driven' activities, such as absolute or conditional beliefs agents get after an event takes place that triggers a *belief revision* (van Benthem 2007a; Baltag and Smets 2006), with formulas like:

$$[\Uparrow\varphi]B_i\psi \quad \text{after receiving 'soft information' that } \varphi, \text{ agent } i \text{ believes that } \psi. \tag{3.2}$$

3.1.3 Preference Change, and Beyond

Once on this road, since rational action is about choosing on the basis of information and preference, it was only a matter of time before dynamic *preference change* and its triggering events became a topic of investigation. This paper will report on some of these developments. And logical dynamics does not even stop here. In principle, any static aspect of agency or language use studied in the existing logical tradition can be 'dynamified', including shifts in temporal perspective, group standing, etc. (cf. van Benthem et al. 1997). One issue which then arises in the logical study of agency is how all these separate dynamifications hang together. Can we really just look at events that produce knowledge, belief, or preference separately, and put them together compositionally? Or is there some deeper conceptual entanglement between these notions calling for more delicate formal constructions? All these issues will be discussed for the case of preference below.

3.1.4 Overview

This paper is mainly based on some recent publications in the Amsterdam environment over the last 3 years. Indeed, 'dynamics' presupposes an account of 'statics', and hence we first give a brief survey of preference logic in a simple modal format using binary comparison relations between possible worlds – on the principle that 'small is beautiful'. We also describe a recent alternative approach, where world preferences are generated from criteria or constraints. We show how to dynamify both views by adding explicit events that trigger preference change in the models, and we sketch how the resulting systems connect. Next, we discuss some entanglements between preference, knowledge and belief, and what this means for combined dynamic logics. On top of this, we also show how more delicate aspects of preference should be incorporated, such as its striking 'ceteris paribus' character, which was already central in von Wright (1963). Finally, we relate our considerations to social choice theory and game theory.

3.2 Modal Logic of Betterness

Preference is a multi-faceted notion: we can prefer one object, or one situation, over another – but preference can also be directed toward kinds of objects or generic types of situation, often defined by propositions. Both perspectives make sense, and a bona fide 'preference logic' should do justice to all of them eventually. We start with a simple scenario on the object/world side, leaving other options for later.

3.2.1 Basic Models

In this paper, we start with a very simple setting. *Modal models* $\boldsymbol{M} = (W, \leq, V)$ consist of a set of worlds W (but they really stand for any sort of objects that are subject to evaluation and comparison), a '*betterness*' *relation* $\leq$ between worlds (where $a \leq b$ reads 'b is at least as good as a'), and a valuation V for proposition letters at worlds (or, for unary properties of objects). In principle, the comparison relation may be different for different agents, but in what follows, we will suppress agent subscripts $\leq_i$ whenever possible for greater readability. Also, we use the artificial term 'betterness' to stress that this is an abstract comparison relation, making no claim yet concerning the natural rendering of the intuitive term 'preference', about which some people hold passionate proprietary opinions. Still, this semantics is entirely natural and concrete. Just think of decision theory, where worlds (standing for outcomes of actions) are compared as to utility, or game theory, where end nodes of a game tree (standing for final histories of the game) are related by preference relations for the different players. In other words, our simple modal models represent a widespread use of the term 'preference' in science.

3.2.2 Digression: Plausibility

Very similar models have been widely used to model another notion, viz. 'relative plausibility' as judged by an agent. This happens in the semantics of belief and doxastic conditionals, where beliefs are those propositions that are true in all most plausible relevant worlds – and various kinds of plausibility models are also crucial to the best-known semantics for belief revision. While preference is not the same as plausibility (except for very wishful thinkers), this formal analogy has proven quite helpful as a source of ideas and transfer of results across the two fields.[1] We will return to the issue of more genuine conceptual 'entanglements' between preference and belief later on.

3.2.3 Modal Languages

Over our base models, we can interpret a standard modal language, and see which natural notions and patterns of reasoning can be defined in it. In particular, a modal assertion like $<\leq> \varphi$ will make the following 'local' assertion at a world w:

$$\boldsymbol{M}, w \models <\leq> \varphi \text{ iff there exists a } v \geq w \text{ with } \boldsymbol{M}, v \models \varphi \tag{3.3}$$

i.e., there is a world v at least as good as w which satisfies φ. In combination with other operators, this simple formalism can express many natural notions concerning rational preference-driven action. For instance, consider finite game trees, which are natural models for a dynamic logic of atomic actions (players' moves) and unary predicates indicating players' turns at intermediate nodes and their utility values at end nodes (van Benthem 2002). Van Benthem, van Otterloo and Roy (2006) show how the *backward induction solution* of a finite game[2] is as the unique binary relation *bi* on the game tree satisfying the following modal preference-action law:

$$[bi^*](end \rightarrow \varphi) \rightarrow [move] < bi^* > (end \,\&\, <\leq> \varphi) \tag{3.4}$$

Here *end* is a proposition letter true only at end-points of the game, *move* is the union of all one-step move relations available to players, and * denotes the reflexive–transitive closure of a relation. When unpacked,[3] the formula says that there is no alternative move to the *BI*-move at the current node all of whose outcomes, when

[1] Cf. the analysis of non-monotonic logic via abstract world preference in Shoham (1988).

[2] A famous 'benchmark example' in the logical analysis of games; cf. Harrenstein (2004). Apt and Zvesper (2007) give a logical take on rationality in solution procedures for strategic form games.

[3] Here is a hint – for a complete explanation, cf. the cited paper. The formula says that if the backward induction strategy played from the current node results in end-points with property φ, then every alternative move right now has a backward induction path to some node that sees a φ-node that is at least as good. This will clearly be the case if there is some end-point x reachable by the actual backward induction move, and an end-point y reachable by the alternative move such

playing the *BI*-solution afterwards, would be better. Thus, modal preference logic seems to go well with games.[4]

But there are more examples of its uses. Already Boutilier (1994) observed how such a simple modal language can also define conditional assertions, normally studied per se as a complex new binary modality (Lewis 1973), and how one can then analyze their logic in standard terms.[5] For instance, in modal models with finite pre-orders (see below), the standard truth definition of a conditional $A \Rightarrow B$ reads as 'B is true in all maximal A-worlds' – and this clause can be written as the following modal combination:

$$\Box\,(A \rightarrow <\leq>(A \,\&\, [\leq](A \rightarrow B))), \text{ with } \Box \text{ some appropriate universal modality.}$$

This formula may look complex at first, but it says something which is easily visualized: every A-world has an A-world 'above' it in the ordering such that B holds here, and also in every A-world that is still better.[6] Now the point is that, in this way, the usual inferential behavior of the conditional, including its well-known non-monotonic features, can be completely understood via the base logic for the unary modalities, say, as a sub-theory of modal $S4$. Moreover, the modal language easily defines variant notions whose introduction seems a big deal in conditional logic, such as existential versions saying that each A-world sees *at least one* maximal $A-$world which is B. Of course, explicit separate axiomatizations of these defined notions retain an independent interest: but we now see the whole picture.[7]

3.2.4 Constraints on Betterness Orders

Which properties should a betterness relation have? Many authors like to work with *total orders*, satisfying reflexivity, transitivity, and connectedness. This is also common practice in decision theory and game theory, since these properties are enforced by the desired numerical representation of agents' utilities. But if we look at the logical literature on preference or plausibility, things are less clear, and properties have been under debate ever since the pioneering study by Halldén (1957). For example, transitivity has been extensively criticized as a constraint on intuitive preference (Hansson 2001). And in conditional logic, Lewis' use of totality is often abandoned

---

that $y \leq x$. (There is even an equivalence, by a modal correspondence argument.) Equivalently, this forbids that some alternative would be strictly better than the backward induction move.

[4] This, and also the following examples are somewhat remarkable, because there has been a widespread prejudice that modal logic is not very suitable to formalizing preference reasoning.

[5] This innovative move is yet to become common knowledge in the logical literature.

[6] On finite models, this is equivalent to demanding that all maximal A-worlds are B-worlds.

[7] Axiomatizing such defined notions per se may be seen as the point of the usual completeness theorems in conditional logic. Also, Halpern (1997) axiomatized a defined notion of preference of this existential sort.

in favor of just *pre-orders*, satisfying the conditions of reflexivity and transitivity, while acknowledging *four* intuitively irreducible basic relations between worlds:

$$w \leq v, \neg v \leq w \text{ (often written as } w < v) \; w \text{ strictly precedes } v \tag{3.6}$$

$$v \leq w, \neg w \leq v \text{ (often written as } v < w) \; v \text{ strictly precedes } w \tag{3.7}$$

$$w \leq v, v \leq w \text{ (sometimes written as } w \sim v) \; w, v \text{ are indifferent} \tag{3.8}$$

$$\neg w \leq v, \neg v \leq w \text{ (sometimes written as } w \# v) \; w, v \text{ are incomparable} \tag{3.9}$$

We feel this pleads for having a large class of models, noting the extra modal principles enforced through frame correspondence *if* we make the relation satisfy extra constraints.[8] The point of a logical analysis is to impose structure where needed, but also, to identify the 'degrees of freedom' where parameters are to be set in a somewhat loose intuitive notion.

3.2.5 Further Relations?

Finally, we note that there may be a case for having two independent betterness relations in models: a weak order $w \leq v$ for 'at least as good', and a strict order $w < v$ for 'better', defined as $w \leq v \,\&\, \neg v \leq w$. Van Benthem, Girard and Roy (2007) axiomatize the logic of this extended language, with separate modalities for the weak and strict betterness relations, plus elegant principles for their interplay.

For more on the austere modal framework of this section and its unifying power, cf. the dissertation by Girard (2008), who shows, drawing upon much more of the relevant literature than we have discussed here, that our basic 'order logic' is a wide-ranging pilot environment for studying essential patterns in reasoning with preference and belief.[9]

3.3 Defining Global Propositional Preference

As we have said, a betterness relation need not yet capture what we mean by agents' preferences in a more colloquial sense. Indeed, many authors consider 'preference' really a relation between propositions, von Wright (1963) being a famous example.

[8] Some people feel a relation 'is' only a preference relation when we impose constraints like transitivity. But this seems a category mistake. A formal relation in a model is just a mathematical object, though it may come to *stand for* a preference in a context of modeling, which requires some scenario attaching the formal model to some reality being described. Moreover, given several decades of research on preference relations, it seems highly unlikely that there is any stable base se of constraints: preference might be more of a 'family notion'.

[9] We have not even exhausted all approaches cooking in Amsterdam right now. For another kind of modal preference logic in games, including a 'normality' operator, see Apt and Zvesper (2007).

These differences seem largely terminological, which is precisely why debates are often bitter.[10]

3.3.1 Set Lifting

Technically, defining preferences between propositions calls for a comparison of sets of worlds. For a given relation $\leq$ among worlds, this may be achieved by *lifting*. One ubiquitous proposal in relation lifting, also elsewhere, is the $\forall\exists$ stipulation that

$$\text{a set } Y \text{ is preferred to a set } X \text{ if } \forall x \in X\, \exists y \in Y : x \leq y. \tag{3.10}$$

As we said, this was axiomatized by Halpern (1997). But alternatives are possible. Van Benthem, Girard and Roy (2007) analyze von Wright's view as the $\forall\forall$ quantifier stipulation that

$$\text{a set } Y \text{ is preferred to a set } X \text{ if } \forall x \in X\, \forall y \in Y : x \leq y, \tag{3.11}$$

and provide a complete logic. And still further combinations occur. Liu (2008) provides a brief history of further proposals for relation lifting in various fields (decision theory, philosophy, computer science), but no consensus on one canonical notion of preference seems to have ever emerged. This may be a feature, rather than a bug. Preference as a comparison relation between propositions may turn out different depending on the scenario. For instance, in a *game*, when comparing sets of outcomes that can be reached by selecting available moves, players may have different options. One would indeed say that we prefer a set whose minimum utility value exceeds the maximum of another (this is like the $\forall\forall$ reading) – but it would also be quite reasonable to say that the maximum of one set exceeds the maximum of the other, which would be rather like the $\forall\exists$ reading.

3.3.2 Extended Modal Logics

The main insight from the current modal literature on preference is twofold. First, many different liftings are definable in our modal base logic extended with a universal modality $U\varphi$: 'φ is true in all worlds'. This standard feature from 'hybrid logic' gives some additional expressive power without great cost in the modal model theory and the computational complexity of valid consequence. For instance, the $\forall\exists$ reading of preference is expressed as follows, with formulas standing for definable sets of worlds:

$$U(\varphi \rightarrow <\leq>\psi). \tag{3.12}$$

[10] Compare William James' famous squirrel going 'round' the tree (or not...): cf. James (1907).

To see this, think of φ as defining the above set X, and of ψ as defining the above set Y. In what follows, we will use the notation $P\varphi\psi$ for such lifted propositional preferences.

Of course, eventually, one can also use stronger formalisms for describing preferences, such as first-order logic (cf. Suppes 1957), but this is just the ordinary balance in logic between finding illuminating formalizations of key notions and argument patterns, and the quest for formalisms combining optimal expressivity with computational ease.[11] We have nothing against using richer languages, but modal logic is an attractive first level to start.[12]

3.4 Dynamics of Evaluation Change

But now for preference change! A modal model describes a current evaluation pattern for worlds, as seen by one or more agents. But the reality is that these patterns are not stable. Things can happen that make us *change* these evaluations of worlds. This dynamic idea has been in the air for quite a while now.[13] In particular, van Benthem van, Eijck and Frolova (1993) already proposed a first system for 'changing preferences', as triggered by various actions that can be defined in a dynamic logic. One of their examples was an 'upgrade event' *#(A)* which makes the proposition A 'more important' in the current model by removing all betterness arrows running from A-worlds to better $\neg A$-worlds.[14] In the same period, Boutilier and Goldszmidt (1993) described a dynamic semantics for conditionals $A \Rightarrow B$, in terms of actions which produce a minimal change in a given Lewis-style world comparison relation so as to *make* all 'best' A-worlds in the new pattern B-worlds. This idea was developed much more systematically in Veltman (1996) on the logical dynamics of default reasoning,[15] and subsequent publications such as Tan and van der Torre (1999) on deontic reasoning and the dynamics of changing obligations that lies behind it. In particular, the systems to be discussed in this paper may be traced back to Zarnic (1999) on practical reasoning, which analyzed actions '$FIAT\varphi$' for factual assertions φ as changes in a comparison relation making the

[11] For more on this essential balance between expressive power of a logic and the computational complexity of its notion of validity, cf. van Benthem and Blackburn (2006).

[12] This may be a good point to also acknowledge the long earlier line of work on modal preference logics by van der Torre and various co-authors, cf. e.g., Tan & van der Torre 1999.

[13] We only review one strand here: cf. again Hansson (1995) for a different point of entry.

[14] One motivation given was as a semantics for a 'weak command' in favor of A.

[15] Veltman insists that the *meaning* of conditionals has this dynamic character, making logical formulas 'implicitly dynamic'. Most work that we are reporting on has 'explicit dynamics', and assumes the traditional static meanings for logical formulas, while using these in explicit triggers for dynamic actions which change models. In other words, one can do 'logical dynamics' without committing to 'update semantics' – and vice versa. Stated differently, we are doing conceptual analysis and rational (re-)construction in Carnap's sense, rather than analysis of 'the meaning' of preference or belief statements (if there *is* such a unique natural language meaning at all).

φ–worlds 'best'. Next, again in deontic logic, Yamada (2006) proposed analyzing acceptance of 'commands' as relation changers, and provided some complete logics in the dynamic–epistemic style.

Of course, realistic preference change has many more features than those mentioned here, which will only come to light on a deeper analysis of agents (cf. Lang and van der Torre 2008). Moreover, various formal proposals already exist (cf. Hansson 1995). But in the remainder of this paper, we concentrate merely on logical methodology of the sort found in the Dutch Lowlands.

3.5 A Basic Dynamic Preference Logic

How does a dynamic logic of preference change work? We present some basic features from van Benthem and Liu (2006), starting with about the simplest scenario.

3.5.1 Dynamic Logic of 'Suggestions'

Betterness models will be as before, and so is the modal base language, with modalities $< \leq >$ and U. But the syntax now adds a feature, borrowed from dynamic logic of programs in computer science. For each formula of the language, we add a model-changing action $\#(\varphi)$ of 'suggestion',[16] defined as follows:

For each model $\boldsymbol{M}, w$, the model $\boldsymbol{M}\#\varphi, w$ is $\boldsymbol{M}, w$ with the new relation

$$\leq' = \leq - \{(x, y) | \boldsymbol{M}, x \models \varphi \ \& \boldsymbol{M}, y \models \neg\varphi\}. \tag{3.13}$$

Note that this model change event is a function on models, providing unique values for each $\boldsymbol{M}, w$.

Next, we enrich the formal language by action modalities as follows[17]:

$$\boldsymbol{M}, w \models [\#(\varphi)]\,\psi \text{ iff } \boldsymbol{M}\#\varphi, w \models \psi \tag{3.14}$$

These allow us to talk about what agents will prefer after their comparison relation has changed. For instance, if you tell me to drink Mexican rather than Bavarian beer, and I accept this recommendation in my evaluation of possible worlds, then I now come to prefer Mexican over Bavarian beer, even if I did not do so before.

Now, as in dynamic–epistemic logic, the heart of the dynamic analysis consists in finding the 'recursion equation' explaining when a preference obtains after an action, in so far as the language can express it. Here is the relevant valid principle

[16] This is of course just an informal reading, not a full-fledged analysis of 'suggestion'.

[17] Here the syntax is recursive: the formula φ may itself contain dynamic modalities.

for suggestions, whose two cases can be seen to follow the above definition of the above model change:

$$<\#(\varphi)> <\leq> \psi \leftrightarrow (\neg \varphi \,\&\, <\leq> <\#(\varphi)>\psi) \vee (\varphi \,\&\, <\leq> (\varphi \,\&\, <\#(\varphi)>\psi) \tag{3.15}$$

Theorem. *The dynamic logic of preference change under suggestions is axiomatized completely by the static modal logic of the underlying model class plus the following equivalences for the dynamic modality:*

$[\#(\varphi)]\, p \leftrightarrow p$, where p is an atomic sentence
$[\#(\varphi)] \neg \psi \leftrightarrow \neg[\#(\varphi)]\, \psi$
$[\#(\varphi)]\, (\psi \& \chi) \leftrightarrow [\#(\varphi)]\psi \,\&[\#(\varphi)]\, \chi$
$[\#(\varphi)]\, U\psi \leftrightarrow U\, [\#(\varphi)]\, \psi$
$[\#(\varphi)] <\leq> \psi \leftrightarrow (\neg \varphi \,\&\, <\leq> [\#(\varphi)]\, \psi) \vee ((\varphi \,\&\, <\leq> (\varphi \,\&\, [\#(\varphi)]\, \psi)$.

Proof. These axioms express the following semantic facts, respectively: upgrade does not change atomic facts, upgrade is a function, upgrade is a normal modality, upgrade does not change the domain of worlds of the model, and upgrade follows the definition of suggestion as explained earlier. Applied inside out, these laws reduce any valid formula to an equivalent one not containing any dynamic modalities, for which the given base logic is already complete by assumption.[18] ■

This logic automatically gives us a dynamic logic of upgraded propositional preferences. For instance, we can compute how ∀∃–type preferences $P\psi\chi$ arise:

$$\begin{aligned}
&[\#(\varphi)]\, P\psi\chi \leftrightarrow [\#(\varphi)]\, U(\psi \rightarrow <\leq>\chi) \leftrightarrow \\
&U\, [\#(\varphi)]\, (\psi \rightarrow <\leq>\chi) \leftrightarrow U([\#(\varphi)]\, \psi \rightarrow [\#(\varphi)]\, <\leq>\chi) \leftrightarrow \\
&U([\#(\varphi)]\, \psi \rightarrow (\neg \varphi \,\&\, <\leq> [\#(\varphi)]\chi) \vee ((\varphi \,\&\, <\leq> (\varphi \,\&[\#(\varphi)]\, \chi) \leftrightarrow \\
&P([\#(\varphi)]\, \psi \,\&\, \neg \varphi)\, [\#(\varphi)]\chi \,\&\, P([\#(\varphi)]\psi \,\&\, \varphi)\, (\varphi \,\&\, [\#(\varphi)]\chi).
\end{aligned} \tag{3.16}$$

3.5.2 General Relation Transformers

But this is still just a 'trial run' for one particular kind of preference change. Van Benthem and Liu (2007) also study other relation transformers.

[18] This reductive analysis shows that the process of preference can be analyzed compositionally. Moreover, it shows that the base language was well-designed, in 'expressive harmony' with the dynamic superstructure. Even so, the real dynamic account of *preference change* is of course in *the recursive procedure itself*, and it lies only hidden implicitly in the base language.

For instance, let $\Uparrow(\varphi)$ be the relation change which makes all φ–worlds better than all $\neg\varphi$-worlds, while keeping the old order inside these zones. In preference terms, this makes φ the 'most desirable good', while in terms of belief revision (van Benthem 2007a), it is a piece of 'soft information' making the φ–worlds the most plausible ones – though still leaving a loophole for $\neg\varphi$'*s* perhaps being true. Again, we can find a complete recursion axiom for this notion, this time as follows, using an 'existential modality' E[19]:

$$[\Uparrow(\varphi)] <\leq>\psi \leftrightarrow (\neg\varphi \,\&\, <\leq>[\Uparrow(\varphi)]\psi)) \vee (\varphi \,\&\, <\leq>(\varphi \,\&\, [\Uparrow(\varphi)]\psi)) \vee (\neg\varphi \,\&\, E(\varphi \,\&\, [\Uparrow(\varphi)]\psi)) \tag{3.17}$$

But in principle, there are many triggers for betterness change, depending on how people adjust to what others claim, command, etc. Thus, it is hard to specify just a small set of changes, with logic serving as an arbiter of how one should respond to them. The task of a dynamic logic of preference is rather providing the appropriate generality, and spotting where some 'trigger' is needed as input to the update.[20]

Here is one way of achieving parametrization of preference change. The new betterness relations in our examples are *definable* from the old ones in the following straightforward syntactic '*PDL* program format', involving *tests?* φ for the truth of propositions φ, and binary *sequential composition*; and $union\cup$ of relations. Let R be the current relation:

$$\#(\varphi)(R) = (?\varphi; R; ?\varphi) \cup (?\neg\varphi; R; ?\neg\varphi) \cup (?\neg\varphi; R; ?\varphi) \tag{3.18}$$

$$\Uparrow(\varphi)(R) = (?\varphi; R; ?\varphi) \cup (?\neg\varphi; R; ?\neg\varphi) \cup (?\neg\varphi; T; ?\varphi) \tag{3.19}$$

where 'T' is the universal relation in the model.

Note that the former definition can only go to a sub-relation of the current one, while the second may add new links as well. Both fall under the following result:

Theorem. *Any relation transformer* τ *with a program definition in the PDL format has a complete reduction axiom, and the latter can be computed effectively from* τ*'s definition.*

The proof is a simple recursive recipe, viewing the definitions basically as 'substitutions' of new relations for old. There are also other ways of achieving generality, e.g., in terms of 'event models' (see Section 3.10 below), but the program method, too, is powerful.[21]

[19] Van Benthem (2007a) uses this axiom to analyze agents' *conditional beliefs* after receiving soft information, with a recursion based on the definition of such beliefs in our modal base language.

[20] Many people have the mistaken belief that this 'plurality' is reprehensible wantonness, whereas localizing the proper *degrees of freedom* for an agent is a precisely a key task for logical analysis.

[21] Van Eijck (2008) uses this technique for belief revision, linking up with 'factual change' in *DEL* as treated in van Benthem, van Eijck and Kooi (2006).

3.5.3 Constraints on Betterness Ordering Once More

While adding a dynamic superstructure to an existing modal logic seems a somewhat 'conservative' enterprise of mere addition, there are several points where matters can be more interesting. One is that, if a static base language is to have enough power for 'pre-encoding' the effects of dynamic changes, it must have the right expressiveness. A good example are the static conditional beliefs needed to pre-encode effects of belief revision, or the 'conditional common knowledge' of van Benthem, van Eijck and Kooi (2006) needed for pre-encoding group knowledge that arises after public announcement. Indeed, some basic logical notions seem to have just this 'forward-looking' character. Such issues of language design may be relevant to preference logic once we study group preferences, but we have not encountered them yet.

But another issue has been noted in van Benthem and Liu (2007). Suppose that our current betterness order satisfies some relational constraints, what guarantees that its transformed version will still satisfy these same constraints? For instance, it is easy to see that the above suggestions take pre-orders to pre-orders, but they can destroy the *totality* of a betterness order. Liu (2008, Chapter 4) analyzes this further, but we have no general results yet. There is an interesting debate here. Some people see this potential loss of basic order properties as a basic drawback of the relation transformer approach. But we feel that the situation is exactly the other way around. The fact that some natural relation transformers break certain relational constraints on preference shows how 'fragile' these constraints really are, and they provide natural scenarios for counter-examples.

3.5.4 Coda: What Was the Case Versus What Should Become the Case

It is tempting to read instructions like $\Uparrow(\varphi)$ as '*see to it that* you come to prefer, or believe, that φ'. This is a forward-oriented view of dynamics: one should make some minimal change resulting in the truth of some stated 'postcondition'. But this is not really the spirit of dynamic–epistemic logic, which rather lets events tell us the 'preconditions' of their occurrence. The two views clash, e.g., in deontic logic, when a command says that you must make sure some proposition becomes true without telling you how. In principle, our approach is 'constructive': triggers in the logic must tell us exactly how the model is to be changed. For the other view, temporal logics (Belnap et al. 2001; van Benthem and Pacuit 2006) may be the better format, where the model already gives the possible future histories. Such models may be seen as enriching dynamic logics with protocols constraining the runs of the relevant informational or preferential process.

3.6 Alternative: Constraint-Based Preference

So far, we have followed the beaten modal path, starting from an ordering of worlds, and deriving notions of preference that apply to propositions, definable in our languages. But there is also another approach to preference, conceptually equally attractive, which works from the opposite direction. Object comparisons are often made on the basis of *criteria*, and then derived from the way in which we apply these criteria, and prioritize between them. For instance, cars may be compared as to price, safety, and comfort, in some order of importance. In that case, the criteria are primary, and the object or world order is derived. This framework, too, occurs in many scientific settings, including philosophy and economics, with various connections made between the two fields by Rott (2001). Another example of its descriptive power is 'Optimality Theory' in linguistics and general cognitive science (Prince and Smolensky 1993; Smolensky 2006).[22]

3.6.1 First-Order Priority Logic

A recent logical formalization of this approach to preference was given by de Jongh and Liu (2006). In the simplest case, one starts from a finite sequence $\boldsymbol{P}$ of propositions, or properties, and then orders objects as follows:

$$\begin{aligned} &x < y \text{ iff } x, y \text{ differ in at least one property in } \boldsymbol{P}, \text{ and} \\ &\text{the first } P \in \boldsymbol{P} \text{ where this happens is one with } Py, \neg Px. \end{aligned} \tag{3.20}$$

This is really a special case of the well-known method of lexicographic ordering, if we view each property $\mathrm{P} \in \boldsymbol{P}$ as inducing the following simple object order[23]:

$$x \leq^P y \text{ iff } (Py \rightarrow Px). \tag{3.21}$$

De Jongh and Liu give a first-order toy language for describing these induced preferences between objects. They also prove a matching representation result for object or world models:

Theorem. *The orders produced via linear 'priority sequences' are precisely the total ones, which satisfy reflexivity, transitivity, and quasi-linearity:* $\forall xyz\,(x \leq y \rightarrow (x \leq z \vee z \leq y))$.

[22] By the way, note that a priority order among propositions need not be a preference relation. I do not 'prefer' safety of my vehicle to sleek design, I just consider it more essential.

[23] We will be free-wheeling in what follows between weak versions $\leq$ and strict ones $<$; but everything we say applies equally well to both versions and their modal axiomatizations.

Liu (2008) discusses this situation further, and notes that the literature has many other ways of defining object order from property orders, which can be studied in similar ways. This diversity may be compared with that for 'lifting' object order to world order before.

3.6.2 Dynamics

Again, this style of analysis suggests an obvious engine for preference change. This time, it is the priority order and set of relevant properties which can change, thereby inducing a change in the defined object order. A new criterion may become relevant, or a criterion may lose its former importance. De Jongh and Liu (2007) study four main operations: *permuting* properties in a priority sequence, *prefixing* a new property, *postfixing* a new property, and *inserting* a property at some specified position. Together, these allow for any manipulation of finite sequences. Moreover, they lead to complete dynamic logics for the changed derived object-level preferences after such changes have taken place first at the level of the prioritized properties. The format is borrowed from the earlier modal one, and therefore, we do not repeat the precise results here. What all this does show is that the style of dynamification in earlier sections also works for first-order logics, making our modal setting a convenience, rather than a straightjacket.

One interesting thing is that the priority dynamics has its own intuitions, different from the account of 'suggestions' or 'commands' we had before. For instance, Girard (2008) re-interprets it as a sort of *agenda* for investigation, determining what is more important than what. He then links the dynamics of 'agenda change' to issues in the philosophy of science, where 'research programs' serve as agendas that keep changing over time.

3.6.3 Two-Level Connections

The two approaches so far may be viewed as complementary.[24] One either starts from a primitive betterness relation between worlds and then lifts it to obtain propositional preference orders, or one starts from a primitive 'importance order' of propositions, and then derives a world order. It is of obvious interest to compare, and perhaps combine the two perspectives, and Liu (2008) has an extensive discussion. To do so, she considers *two-level structures* $(W, \leq, \boldsymbol{P}, <)$ having both worlds with a betterness order $\leq$ and a set of 'important propositions' with a primitive priority order $<$ (cf. Fig. 3.1):

[24] There are obvious connections here with the duality in belief revision between working with a basic world order, or a primitive '*entrenchment order*' of propositions; cf. the excellent survey in Gärdenfors and Rott (1995); but we do not pursue this analogy here.

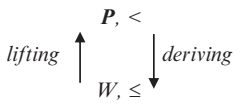

Fig. 3.1 Two-level preference structures

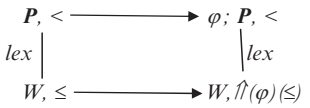

Fig. 3.2 Lexicographic derivation of object order

This picture immediately suggests a number of questions, many of them still unresolved. For example, structurally, what happens when we derive a betterness order form a priority order, and then lift it again? And what happens vice versa?[25] And in terms of languages, what happens when we treat the propositions in $\boldsymbol{P}$ as distinguished propositional constants in a modal language, and try to relate modal betterness logic with modal constraint logic? We have no general answers here, but Liu (2008) does state elegant correspondences between the dynamics at the two levels. In particular, she shows that prefixing of propositions φ to a current priority sequence $\boldsymbol{P}$ has the same effect as the earlier relation transformer $\Uparrow(\varphi)$. More precisely, writing the lexicographic derivation of object order as a function *lex*, the following identity holds, making the following diagram commute (Fig. 3.2):

$$lex(\varphi; \boldsymbol{P}) = \Uparrow(\varphi)(lex(\boldsymbol{P})) \tag{3.22}$$

Again, the general theory of inducing dynamics from one level to another seems open. There also seems to be room here for a more general calculus of natural operations on priority sequences, called 'agenda algebra' in Girard (2008).[26]

3.7 Further Aspects of Preference: Ceteris Paribus Logic

All our logics so far, whether betterness- or priority-based, described pure preferences. But in reality, preferences usually have a defeasible character: they hold only *ceteris paribus*, in von Wright's terminology. Van Benthem, Girard and Roy (2007) discuss this feature, and describe what needs to be changed in our modal approach to make it a more realistic account of reasoning with preferences.

[25] Nevertheless, as we said before, a priority order is not necessarily a preference order.

[26] For instance, for each set of properties, there is a set of *disjoint* properties generating the same object order. Finding the latter effectively is a matter of merging Boolean normal form principles with some preference logic. A few first principles are found in the cited references.

3.7.1 Normality Versus Equality

First, the term 'ceteris paribus', though widely used, has no unambiguous meaning. In fact, one can distinguish two main views. In many scenarios, the *normality sense* says that we only make the preference comparison 'under normal circumstances'. I prefer beer over wine, but not when dining at the Paris Ritz. This may modeled by the 'normal' or most plausible worlds' of our current model. These worlds are singled out, either by some explicit description N, or just as the most plausible worlds in some doxastic plausibility order. In the former scenario, our earlier logic still suffices. We could express a global preference $P\varphi\psi$ in this normality sense as

$$P(N\&\varphi)(N\&\psi). \tag{3.23}$$

But this approach by explicit definition of normal worlds will not work in general, and then we must use models with both betterness and plausibility orders, as in Lang, van der Torre and Weydert (2003), with some matching combined logic of preference and belief. We will return to this issue of what may be called 'entanglement' in the next section.

For now, we note that there is also another *equality sense* of 'ceteris paribus': indeed, the one favored by von Wright. In this sense, a preference statement is made globally, though under the proviso that certain propositions do not change their truth values. For instance, someone who generally prefers work over vacation, might still be said to prefer night over day with work/vacation 'frozen' in a 'ceteris paribus', even though there are vacation days that she would prefer to work nights. More precisely, for von Wright, a ceteris paribus preference for φ over ψ with respect to some proposition A means that

$$\begin{aligned}&\text{both (i) among the } A\text{–worlds I prefer } \varphi \text{ over } \psi,\\ &\text{and (ii) among the } \neg A - \text{ worlds I prefer } \varphi \text{ over } \psi.\end{aligned} \tag{3.24}$$

Thus, cross-comparisons between the A and $\neg A$ worlds are irrelevant to the truth of the preference.[27] For the case of more relevant propositions $\boldsymbol{A}$, one looks at the equivalence classes of worlds under the relation $\equiv_A$ of 'sharing the same truth values on the $\boldsymbol{A}$'s'. von Wright himself proposed a particular set of relevant propositions $\boldsymbol{A}$ to be kept 'constant', viz. all the proposition letters of the language that do not occur in the two formulas φ, ψ being compared in a preference statement $P\varphi\psi$. His preference logic has explicit rules of reasoning expressing this feature (von Wright 1963).

This scenario is interesting because the same relation $\equiv_A$ has been studied elsewhere as an account of the intuitive notions of 'dependence' and 'independence' among propositions (Doyle and Wellman 1994). It also occurs in the semantics of questions and answers in natural language (ten Cate and Shan 2002), and in

[27] This is a conjunction of two 'normality' readings: one with $N = A$, and one with $N = \neg A$.

treatments of supervenience and dependence in philosophy. Thus there is some logical interest to formalizing this.[28]

3.7.2 Equality-Based Ceteris Paribus Preference Logic

Van Benthem, Girard and Roy (2007) make equality-based ceteris paribus preferences an explicit part of the language, making reasoners specify explicitly which propositions are to be 'frozen' in their comparisons. They give a modal logic *CPL* extending basic preference logic with operators

$$\boldsymbol{M}, s \models [\Gamma]\varphi \text{ iff } \boldsymbol{M}, t \models \varphi \text{ for all } t \text{ with } s \equiv_\Gamma t, \tag{3.25}$$

$$\boldsymbol{M}, s \models [\Gamma]^{\leq}\varphi \text{ iff } \boldsymbol{M}, t \models \varphi \text{ for all } t \text{ with } s \equiv_\Gamma t \text{ and } s \leq t, \tag{3.26}$$

$$\boldsymbol{M}, s \models [\Gamma]^{<}\varphi \text{ iff } \boldsymbol{M}, t \models \varphi \text{ for all } t \text{ with } s \equiv_\Gamma t \text{ and } s < t. \tag{3.27}$$

Then a Γ-equality-based ceteris paribus preference $P\varphi\psi$ can be defined as follows:

$$U(\varphi \rightarrow <\Gamma>^{\leq} \psi) \tag{3.28}$$

In practice, the sets Γ are often finite, but the system also allows infinite sets, with even recursion in the definition of the ceteris paribus formulas. For the finite case, we have:

Theorem. *The static logic of CPL is completely axiomatizable.*

Proof. The idea is this. All formulas in the new language have an equivalent formula in the base language thanks to the basic laws for manipulating ceteris paribus riders. The most important one of these tell us how to change the sets Γ:

$$<\Gamma'>^{\leq} \varphi \rightarrow <\Gamma>^{\leq} \varphi \quad \text{if } \Gamma \subseteq \Gamma'$$
$$((\neg)\alpha \,\&\, <\Gamma>^{\leq} ((\neg)\alpha \,\&\, \varphi) \rightarrow <\Gamma \cup \{\alpha\}>^{\leq} \varphi$$

Applying these laws iteratively inside out will remove all ceteris paribus modalities from a formula cases $< \emptyset >^{\leq}$ remain, i.e., ordinary preference modalities from the base system. ■

The main contribution here is an explicit calculus for reasoning with ceteris paribus propositions. This improves over von Wright, where the set Γ is implicit in the context, with some tricky features. For instance, von Wright's account of preference reasoning has no *monotonicity* in the sense that $P\varphi\psi$ implies $P(\varphi \,\&\, \alpha)\psi$, even though this inference seems plausible. The reason is that the extended formula

[28] For more general logics of dependence, cf. van Benthem (1996), Väänänen (2007).

$\varphi \,\&\, \alpha$ *changes* the set of relevant ceteris paribus propositions insidiously, a phenomenon explicit in the indexed modalities of the logic *CPL*, which wears the true monotonicity properties upon its sleeves.

3.7.3 Further Developments

The *CPL* axioms for changing ceteris paribus sets suggest an underlying dynamic process of context change, or in earlier terms, 'agenda change'. Van Benthem, Girard and Roy (2007) also give a *dynamic logic version* of the system, where the 'agenda' is an independent item, which can be extended or simplified – though not all natural operations admit of *DEL*-style recursion axioms. Another source of open problems is the full infinitary version of the system, which is still bisimulation-invariant, but sits somewhere in the landscape of *infinitary modal logics* at some distance from propositional dynamic logic, or other well-behaved calculi. Finally, the connection with *logics of dependence* is intriguing, but not yet understood. For instance, dependence patterns occur typically also in preference reasoning in game theory, our initial example. The authors show that Nash Equilibrium can be defined in their logic, but for this, they use only their local modality looking at worlds (i.e., strategy profiles in the game setting) having the same strategies for the other players as the current world (profile).[29] This seems more like the normality sense of ceteris paribus. The more sweeping equality sense would look at all equivalence classes arising from fixing any strategy profile for the other players, thus moving closer to game-theoretic notions like 'strictly dominated strategies'.

3.8 Entanglement: Preference, Knowledge, and Belief

Now we get to an issue which tends to generate heat among academics. So far, we have analyzed preference per se, as a mere matter of betterness comparison across worlds. But to many people, preference is a deeply *epistemic* or *doxastic* notion, manipulable by changes in beliefs, and subject to introspection. Can we do justice to such intuitions? The standard 'piecewise' approach here would be to add epistemic or doxastic structure to our models, and then define 'real preference' in terms of operator combinations with the earlier modalities for betterness as well as knowledge and belief. Or should the marriage be more intimate? We discuss these issues briefly, following Liu (2008, Chapter 4). It should be noted that they come up in different settings, and, e.g., de Jongh and Liu (2007) make belief-based preference their central notion, providing a complete first-order-style axiomatization. In what follows, we explain the main issues in a modal setting.

[29] As we have said before, there is a flourishing literature on logics providing definitions for basic game-theoretic notions, so it is the ceteris paribus aspect that is of interest here.

3.8.1 First Degree of Entanglement: Combine Separate Operators

Van Benthem and Liu (2007) present a combined system with both knowledge and preference, whose models have both epistemic accessibility relations and a preference order. Their formal language has both betterness modalities $<\leq>$ as before, the auxiliary universal modality, and epistemic *knowledge modalities* $K\varphi$ interpreted as usual as truth of φ in all epistemically accessible worlds. This language can interpret delicate nested operator combinations such as

$$KP\varphi\psi \quad \text{knowing that some global betterness relationship holds,} \tag{3.29}$$

$$PK\varphi K\psi \quad \text{preferring to know certain things over others}^{30}. \tag{3.30}$$

The semantics typically allows for comparisons beyond epistemically accessible worlds. This gives it the option of expressing a sense of 'regret' in which I prefer marching in the Roman Army to being a peaceful academic, even though I know that, alas, the former alternative cannot be. Of course, more realistic (and less romantic) agents will not use this facility provided by the system, and the logic does not force them to.

This language improves on the earlier definition of global preferences $P\varphi\psi$, reading the earlier $U(\varphi \rightarrow <\leq>\psi)$ with a universal modality in epistemic terms:

$$K(\varphi \rightarrow <\leq>\psi). \tag{3.31}$$

3.8.2 Public Announcement Logic

Next, the dynamics in the system will have two forms. There are the betterness changing events we described before, but there are also purely informative events like a public announcement or public observation $!\varphi$ of some φ true right now in the actual world. These are the simplest forms of learning some new piece of 'hard information'. They are treated in the standard format of dynamic–epistemic logic, as a restriction of the current model $\boldsymbol{M}, s$ to its sub-model $\boldsymbol{M}|\varphi, s$ consisting of the worlds satisfying φ in $\boldsymbol{M}$. Again, we extend the language with modalities, this time as follows:

$$\boldsymbol{M}, s \models [!\varphi]\psi \;\text{ iff }\; \textit{if}\, \boldsymbol{M}, s \models \varphi, \textit{then}\, \boldsymbol{M}|\varphi, s \models \psi \tag{3.32}$$

Here the condition expresses the precondition that the new information is true, and hence update is only a partial function. The key recursion principle of the resulting *public announcement logic* (cf. Gerbrandy 1999 van Benthem 2006) is the following law, which describes which knowledge arises from receiving hard information:

[30] This combination raises some tricky issues of intuitive interpretation, which might work better in an epistemic or doxastic temporal logic that can deal with scenarios of investigation.

$$[!\varphi]K_i\psi \leftrightarrow (\varphi \rightarrow K_i(\varphi \rightarrow [!\varphi]\psi)) \tag{3.33}$$

This structure is easily combined with the earlier dynamic logics of preference change. For instance, as a special case we have

Theorem. *The combined logic of public announcement and suggestion consists of all separate principles for these operations plus two 'cross-comparisons' describing betterness after update and knowledge after upgrade:*

$$[!\varphi] <\leq>\psi \leftrightarrow (\varphi \rightarrow <\leq>(\varphi \,\&\, [!\varphi]\psi))$$
$$[\#\varphi]K_i\psi \leftrightarrow K_i[\#\varphi]\psi$$

This logic can handle scenarios with both information and preference changes.

3.8.3 Digression: Upgrade Versus Update

Sometimes, the above even offers alternative descriptions for one story. Take the example from Liu (2008) about buying a house – but similar observations have been made by many authors. I am indifferent about buying one near the park or in town, but now I learn that a freeway will be built near the park, and I come to prefer the house in town. This may be described as a *2-world* model

• 'Buy park house', • 'Buy town house'

with an indifference relation between them, where a 'suggestion' upgrade leaves both worlds, but removes a $\leq$-link, leaving a strictly better town house. But alternatively, one could describe the buying scenario in terms of a *4-world* model with extended options

• 'Park house, no freeway', • 'Park house, freeway'

• 'Town house, no freeway', • 'Town house, freeway'

with betterness relations between them. Then a public announcement 'freeway' removes two worlds to get the model we got before by upgrading. We return to this issue below.

Similar points can be made about belief. Take any complete dynamic logic of belief change as found, say, in van Benthem (2007a) or Baltag and Smets (2006), and merge it with any dynamic logic of preference upgrade. This will then deal with combined notions like 'believing that φ is better', or it 'being better to believe φ'.

3.8.4 Second Degree of Entanglement: New Modalities for Intersections

Still, the expressive power of the merged languages described here may not yet be suitable for getting at the real entanglement of preference and knowledge or belief. An epistemized preference formula $K(\varphi \rightarrow <\leq> \psi)$ (subject to introspection, and knowledge-dependent) refers to ψ-worlds that are better than epistemically accessible φ–worlds, but there is no guarantee that these ψ–worlds are *themselves* epistemically accessible. But in our intuitive reading, for instance, of the normality sense of ceteris paribus preference, we made the betterness comparison *inside* the set of normal worlds (cf. again Lang et al. 2003), and likewise, we may want to make it inside the epistemically accessible worlds.[31]

To describe this, it makes sense to introduce a modal language that can talk about the *intersection* of the epistemic relation $\sim$ and the betterness relation $\leq$.[32] That is,

$$\boldsymbol{M}, s \models <\leq \cap \sim> \varphi \text{ iff there is a } t \text{ with } s \sim t \;\&\; s \leq t \text{ such that } \boldsymbol{M}, t \models \varphi \quad (3.34)$$

Now we can define versions of 'internally epistemized' preference, say, claiming that each epistemically accessible φ-world sees an accessible ψ-world that is at least as good:

$$K(\varphi \rightarrow <\leq \cap \sim> \psi) \quad (3.35)$$

This richer logic is no longer bisimulation-invariant, but it is not much more complex than the earlier one. Liu 2008 notes how it supports exactly the same recursive style of dynamic analysis as before. In particular, the following law is valid:

$$\begin{aligned} &< \#(\varphi) > <\leq \cap \sim> \psi \leftrightarrow (\neg\varphi \;\&\; <\leq \cap \sim> < \#(\varphi) > \psi) \vee \\ &(\varphi \;\&\; <\leq \cap \sim> (\varphi \;\&\; <\#(\varphi)> \psi) \end{aligned} \quad (3.36)$$

Again, completely similar points hold for belief instead of knowledge, using intersection modalities with respect to betterness and plausibility relations between worlds.[33] Dynamic informational actions then include both announcements of hard information and various sorts of plausibility-changing 'soft information' that trigger belief revision.

[31] A similar entanglement, this time of epistemic and doxastic structure, is found in the work on belief revision by Baltag and Smets (2006), followed up on by van Eijck (2008).

[32] Intersection is implicit in the truth conditions for the above ceteris paribus logic *CPL*, where betterness became intersected with truth-value equivalence for a formula set Γ.

[33] Note that all issues discussed so far also arise in the *constraint-based* approach of Section 3.6.

3.8.5 Third Degree of Entanglement: Preference and Belief as Duals

Finally, all this piecemeal modal combination might still be too simple and technically driven. Preference and belief may also be taken to be totally inter-definable notions, and much of the literature on the foundations of decision theory (cf. Pacuit and Roy 2006 and the references therein; while Zvesper 2008 connects up with the Lewis–Price discussion on relations between desire and belief) suggests that we can learn a person's beliefs from her preferences, as revealed by her actions[34] – and also vice versa, that we can learn her preferences from her beliefs. We leave the pros and cons of this conceptual connection as an open problem, which actually highlights the broader challenge of relating preference logic to decision theory.

3.8.6 Excursion: Logic and Probability

The entanglement of belief and preference is yet much more intense in probabilistic approaches to preference logic and its dynamics, which also have a longer pedigree than the logical ones discussed here (cf. Hansson 1995; Bradley 2007). For instance, every time we compute an *expected value*, we mix probabilities that can stand in principle for agents' subjective beliefs with utilities that express betterness or preference. And this mixture is much more global than the above entanglement in terms of 'most plausible worlds' intersected with betterness. This interface is worth a separate study, and the 2008 Amsterdam Workshop 'The Dynamics of Preferences and Intentions' (http://staff.science.uva.nl/∼oroy/GLLC15/) gave a first glimpse of what might happen.

3.9 Multi-agent Interaction and Group Preference

As a final topic which I see as central to preference logic, another feature of information dynamics also makes sense for preference, viz. its *multi-agent interactive* character which also involves groups as new agents in their own right. For a start, let us look at the most obvious interactive test-bed for logics of preference and information, making the earlier issues much more concrete, viz. *games*.

[34] Cf. Lewis (1988) for a dissenting (though controversial) view on this Humean theme.

3.9.1 Game Theory, Epistemic Preference Logic, and Backward Induction

Combined epistemic preference logics have already been applied to a variety of issues in games. Harrenstein (2004) used them to define Nash equilibrium, and van Otterloo (2005) has a chapter on preferences of players and how these change when further information becomes available about their 'intentions', i.e., the strategies that they will play from now on. Van Benthem (2007b) discusses the role of 'promises' in games, viewed in a similar way as *public announcements of intentions*, while also discussing related settings where players' preferences (encoded as betterness relations on nodes in the game tree) are not known.[35]

The entanglement of knowledge and belief with betterness and preference becomes quite concrete and vivid in this setting. Consider the well-known game solution procedure of *Backward Induction*. In the following picture, an equilibrium with outcomes (*1, 0*) will be computed by inductive bottom-up reasoning about players' 'rationality' – incidentally, making both hugely worse off than the cooperative outcome (*99, 99*) (cf. Fig. 3.3):

As pointed out by Board (1998), van Benthem (2002), the reasoning behind the standard Nash equilibrium here really rests on deriving *expectations* from the given betterness relations among end nodes, and then choosing moves accordingly. More concretely, there are three worlds, one for each complete history of the game, and the backward induction reasoning creates a plausibility ordering among these, which is actually the same for both players, with the world of (*1, 0*) on top, then that with (*0, 100*) and then that with (*99, 99*). Thus in games, the plausibility relations that we merely stipulate in models for belief revision arise from an underlying analysis connecting belief with preference.

But we also see that this entanglement between belief and preference is not 'absolute'. It depends crucially on assumptions that we make about the *type of agent* involved. One can only predict beliefs from people's preferences by assuming, for instance, that they are *rational* utility-maximizing agents in the sense of decision theory or game theory.[36] Thus, I am not yet convinced that preference and belief are

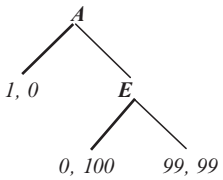

Fig. 3.3 A decision tree

[35] Changes in games may improve their equilibria. For example, in the game that follows, ***E*** might promise that she will not go left, and this public announcement changes the game to one with just the 'right' move for her – and a new equilibrium (*99, 99*) results.

[36] Incidentally, in this setting, it is crucial to make betterness comparisons with worlds that we believe will not happen: it is precisely those worlds that keep the actual prediction 'in place'.

truly dual notions, as a majority view seems to have it. They rather seem separate to me, though they may be connected tightly through making different assumptions on agents. And it would rather be a task for preference logic to sort out what natural assumptions there are, in addition to the ubiquitous 'rationality', and how they may become subject to explicit reasoning.

3.9.2 Preferences and Intentions

Much more sophisticated scenarios are discussed in the dissertation Roy (2008), an extensive logic-inspired study of the role of *intentions* and commitments in decision making and game playing. Rational intentions are based on preferences, but they add further aspects of agents' capabilities and their *plans* for achieving goals, which are beyond our simple preference-based logic frameworks so far.[37] While these richer models are definitely worthwhile, they lie beyond our horizon here.

3.9.3 Preference Merge and Group Modalities

While games still involve interaction between individual agents by themselves, the next obvious step is to introduce *groups* themselves as new collective agents. Indeed, game theorists study coalitions, while in epistemic and doxastic logic, common knowledge or common belief of groups has become a standard notion in understanding stable behavior in communication and interaction.[38] The corresponding issue in preference logic would be how group preferences arise out of individual ones. This issue has also come up in belief revision theory, under the name of 'belief merge' for groups of agents who need to merge their plausibility relations.

A highly sophisticated paradigm for relation merge among many agents is that proposed by Andréka, Ryan and Schobbens (2002). It puts the relations to be merged in an ordered *priority graph* $\boldsymbol{G} = (G, <)$ of indices (which may have multiple occurrences), and sets

$$x \leq_G y \text{ iff for all indices } i \in \boldsymbol{G}, \text{ either } x \leq_i y, \text{or there is some } j > i \text{ in } \boldsymbol{G} \text{ with } x <_i y$$[38]

Girard (2008) and Liu (2008) show how this elegant set-up generalizes (amongst many other things) the priority sequences of de Jongh and Liu (2007) in Section 3.6, as well as the 'agendas' we hinted at in connection with ceteris paribus preference

[37] Cf. Pacuit (2008) on links with work on planning and intention change in temporal logic.

[38] A nice combined logic for actions and preferences of *coalitions* in games is Kurzen (2007).

[38] Thus, either x comes below y, or if not, y 'compensates' for this by doing better on some comparison relation in the set with a higher priority in the graph.

logic (Section 3.7). Andréka, Ryan and Schobbens (2002) prove a number of interesting mathematical results about priority graphs, including their universality as a preference aggregation procedure for hierarchical groups, and a complete algebraic axiomatization. Girard (2008) provides an alternative complete axiomatization in a suitable modal language.

As for *dynamics* in this new two-level perspective, there are some natural operations for changing and combining priority graphs, viz. their *sequential* and *parallel composition*. These lead to an elegant calculus of graph operations and their induced group preference relations. This may be viewed as a compositional logic of group preference, much richer than the set-based approaches which have dominated the literature – and it applies equally well to preference formation as belief merge.

3.9.4 Dynamics of Social Choice

All this points at a junction between preference logic including group preferences and *social choice theory*. This is indeed where things seem to be heading these days. Preference logics with group preferences seem to be the natural counterpart to epistemic logics with various forms of group knowledge, and taken together, they provide a rich account of groups that can learn and form new preferences. Of course, much remains to be understood concerning the fine-structure of informative actions for groups, the ways in which they *deliberate*, and the ways in which agents are subject to preference change. These include at least two processes: (a) adjustment of one's initial preferences through social encounters, and (b) even leaving initial individual preferences intact, joining in the formation of new groups with preferences of their own. The empirical reality of voting procedures, and rules for rational discussion and debate would seem to provide excellent challenges for extended preference logic in this sense.

3.10 Conclusions and Further Issues

We have given an overview of dynamic logics of preference change as being developed in Amsterdam, first for individual agents, and eventually also for groups of agents. Many topics have been suppressed in this sketch,[39] such as the use of *product update* (Baltag and Smets 2006) as a congenial but different methodology, *numerical plausibility and utility change* (dating back to Aucher 2003), and in particular, connections and contrasts with *probability* and decision theory. As to the latter, as we already noted, so far, nothing in our preference logics, 'entangled' or not, matches the role of *expected value* in decision and game theory, where utilities of alternative options are weighed probabilistically. How serious is this limitation?

[39] As well as ILLC Gloriclass fellows Andreas Witzel, Joel Uckelmans, and Cédric Dégrémont.

Does it relegate preference logic, no matter how broad and 'dynamic', to the sidelines forever? We do not know, but we do think that the presentation given here links preference logic in its traditional guise to exciting new developments in logic, computation, belief revision, and social choice theory (cf. Endriss and Lang 2006). And maybe that is quite enough for one paper.

Acknowledgment I wish to thank Fenrong Liu and Jonathan Zvesper for their useful comments.

References

Andréka, H., M. Ryan, and P.-Y. Schobbens. 2002. Operators and Laws for Combining Preference Relations. *Journal of Logic and Computation* 12(1): 13–53.

Apt, K. and J. Zvesper. 2007. Common Beliefs and Public Announcements in Strategic Games with Arbitrary Strategy Sets'. Mimeo, CWI and ILLC, University of Amsterdam.

Aucher, G. 2003. A Combined System for Update Logic and Belief Revision. Master's thesis, ILLC, University of Amsterdam.

Baltag, A. and S. Smets. 2006. Dynamic Belief Revision over Multi-Agent Plausibility Models. In *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory*, eds. G. Bonanno, W. van der Hoek and M. Woolridge. To appear in Texts in Logic and Games, Amsterdam: Amsterdam University Press, 2008.

Baltag, A., H. van Ditmarsch, and L. Moss. 2008. Epistemic Logic and Information Update. To appear in *Handbook of the Philosophy of Information*, eds. P. Adriaans and J. van Benthem. Amsterdam: Elsevier.

Belnap, N., M. Perloff, and M. Xu. 2001. *Facing the Future*. Oxford: Oxford University Press.

Board, O. 1998. Belief Revision and Rationalizability. In Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge, ed. I. Gilboa. San Francisco, CA: Morgan Kaufman, 201–213.

Boutilier, C. 1994. Conditional Logics of Normality: A Modal Approach. *Artificial Intelligence* 68: 87–154.

Boutilier, C. and M. Goldszmidt. 1993. Revision by Conditional Beliefs. In *Proceedings AAAI 11*, Washington, DC: Morgan Kaufmann, 649–654.

Bradley, R. 2007. The Kinematics of Belief and Desire. *Synthese* 156(3): 513–535.

Doyle, J. and M. Wellman. 1994. Representing Preferences as Ceteris Paribus Comparatives. In *Decision-Theoretic Planning: Papers from the 1994 Spring {AAAI} Symposium*, 69–75. Menlo Par, CA: AAAI.

Endriss, U. and J. Lang, eds. 2006. *Proceedings of the 1st International Workshop on Computational Social Choice (COMSOC-2006)*, The Netherlands: ILLC, University of Amsterdam.

de Jongh, D. and F. Liu. 2006. Optimality, Belief, and Preference. In *Proceedings of the Workshop on Rationality and Knowledge*, eds. S. Artemov and R. Parikh, PP-PP. Malaga, Spain: ESSLLI Summer School.

Gärdenfors, P. and H. Rott. 1995. Belief Revision. In *Handbook of Logic in Artificial Intelligence and Logic Programming 4*, eds. D. M. Gabbay, C. J. Hogger and J. A. Robinson, PP-PP. Oxford: Oxford University Press.

Gerbrandy, J. 1999. Bisimulations on Planet Kripke. Dissertation, Institute for Logic, Language, and Computation, University of Amsterdam.

Girard, P. 2008. Modal Logics for Belief and Preference Change. Dissertation, ILLC, University of Amsterdam and Department of Philosophy, Stanford University.

Halldén, S. 1957. *On the Logic of "Better"*. Lund, Sweden: Gleerup.

Halpern, J. 1997. Defining Relative Likelihood in Partially-Ordered Preferential Structure. *Journal of Artificial Intelligence Research* 7: 1–24.

Hansson, S. O. 1995. Changes in Preference. *Theory and Decision* 38: 1–28.
Hansson, S. O. 2001. Preference Logic. In *Handbook of Philosophical Logic IV*, eds. D. Gabbay and F. Guenthner, 319–393. Dordrecht, The Netherlands: Kluwer.
Harrenstein, P. 2004. Logic in Conflict. Dissertation, Institute of Computer Science, University of Utrecht.
James, W. 1907. *Pragmatism: A New Name for Some Old Ways of Thinking*. New York: David McKay.
Kurzen, L. 2007. A Logic for Cooperation, Actions and Preferences. Master's thesis, Institute for Logic, Language and Computation, University of Amsterdam. Paper version presented at KRAMAS 11, Sydney, September 2008.
Lang, J. and L. van der Torre. 2008. From Belief Change to Preference Change. TYPE OF PAPER, IRIT Toulouse and University of Luxemburg.
Lang, J., L. van der Torre, and E. Weydert. 2003. Hidden Uncertainty in the Logical Representation of Desires. In *Proceedings IJCAI XVIII*, 685–690.
Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.
Lewis, D. 1988. Desire as Belief. *Mind* 97: 323–332.
Liu, F. 2008. Changing for the Better: Preference Dynamics and Agent Diversity. Dissertation, Institute for Logic, Language and Computation, University of Amsterdam.
Pacuit, E. 2008. Toward a Dynamic Logic of Intention Revision. Mimeo. Department of Computer Science, Stanford University.
Pacuit, E. and O. Roy. 2006. Preference Based Belief Dynamics. In *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory*, eds. G. Bonanno, W. van der Hoek and M. Woolridge To appear in Texts in Logic and Games, Amsterdam: Amsterdam University Press, 2008.
Prince, A. and P. Smolensky. 1993. Optimality Theory: Constraint Interaction in Generative Grammar. Rutgers University Center for Cognitive Science Technical Report 2.
Rott, H. 2001. *Change, Choice and Inference*. Oxford: Oxford University Press.
Roy, O. 2008. Thinking Before Acting: Intentions, Logic, and Rational Choice. Dissertation, Institute for Logic, Language and Computation, University of Amsterdam.
Shoham, Y. 1988. *Reasoning About Change*. Cambridge, MA: MIT Press.
Smolensky, P. 2006. *The Harmonic Mind*. Cambridge, MA: MIT Press.
Suppes, P. 1957. *Introduction to Logic*. Princeton, NJ/New York: Van Nostrand.
Tan, Y-H. and L. van der Torre. 1999. An Update Semantics for Deontic Reasoning. In *Norms, Logics and Information Systems*, eds. P. McNamara and H. Prakken, 73–90. IOS Press, Amsterdam.
ten Cate, B. and Ch-ch Shan. 2002. The Partition Semantics of Questions, Syntactically. In *Proceedings of the ESSLLI*-2002 Student Session, ed. Malvina Nissim, 255–269.
Väänänen, J. 2007. *Dependence Logic*. Cambridge: Cambridge University Press.
van Benthem, J. 1996. *Exploring Logical Dynamics*. Stanford, CA: CSLI.
van Benthem, J. 2002. Extensive Games as Process Models. *Journal of Logic, Language and Information* 11: 289–313.
van Benthem, J. 2006. One Is a Lonely Number: On the Logic of Communication. In Logic Colloquium '02, eds. Z. Chatzidakis, P. Koepke and W. Pohlers, 96–129. Wellesley, MA: ASL & A.K. Peters.
van Benthem, J. 2007a. Dynamic Logic of Belief Revision. *Journal of Applied Non-Classical Logics* 17: 129–155.
van Benthem, J. 2007b. Rationalizations and Promises in Games. *Philosophical Trends* 'Supplement 2006' on Logic: 1–6. Beijing: Chinese Academy of Social Sciences.
van Benthem, J. 2008. Logic, Rational Agency, and Intelligent Interaction. Research Report, ILLC, University of Amsterdam. To appear in *Proceedings 14th Congress of Logic, Methodology and Philosophy of Science Beijing 2007*, eds. D. Westerståhl et al. London: College Publications.
van Benthem, J. and P. Blackburn. 2006. Modal Logic: A Semantic Perspective. In *Handbook of Modal Logic*, eds. P. Blackburn, J. van Benthem and F. Wolter, 1–84. Amsterdam: Elsevier.
van Benthem, J. and F. Liu. 2007. Dynamic Logics of Preference Upgrade. Journal of *Applied Non-Classical Logics* 17: 157–182.

van Benthem, J. and E. Pacuit. 2006. The Tree of Knowledge in Action. In *Proceedings Advances in Modal Logic*, eds. G. Governatori, I. Hodkinson and Y. Venema. Noosa, Queensland, Australia: College Publication.

van Benthem, J., J. van Eijck, and A. Frolova. 1993. Changing Preferences. Report CS-93–10, Centre for Mathematics and Computer Science, University of Amsterdam.

van Benthem, J., R. Muskens, and A. Visser. 1997. Dynamics. In *Handbook of Logic and Language*, eds. J. van Benthem and A. ter Meulen, 587–648. Amsterdam: Elsevier.

van Benthem, J., J. van Eijck, and B. Kooi. 2006. Logics of Communication and Change. *Information and Computation* 204(11): 1620–1662.

van Benthem, J., S. van Otterloo, and O. Roy. 2006. Preference Logic, Conditionals, and Solution Concepts in Games. In *Modality Matters*, eds. H. Lagerlund, S. Lindström and R. Sliwinski, 61–76. Uppsala, Sweden: University of Uppsala.

van Benthem, J., P. Girard, and O. Roy. 2007. Everything Else Being Equal. A Modal Logic Approach to Ceteris Paribus Preferences. Mimieo, Institute for Logic, Language and Computation, University of Amsterdam. To appear in *Journal of Philosophical Logic*, autumn 2008.

van Ditmarsch, H., W. van der Hoek, and B. Kooi. 2007. *Dynamic-Epistemic Logic*. Berlin: Springer.

van Eijck, J. 2008. Commentary to Bonanno. To appear in *Games and Interaction*, Texts in Logic and Games, eds. K. Apt and R. van Rooij. Amsterdam University Press, Amsterdam.

van Otterloo, S. 2005. A Strategic Analysis of Multi-Agent Protocols. Dissertation, DS-2005-05, ILLC, University of Amsterdam and University of Liverpool.

Veltman, F. 1996. Defaults in Update Semantics. *Journal of Philosophical Logic* 25: 221–261.

von Wright, G. H. 1963. *The Logic of Preference*. Edinburgh: Edinburgh University Press.

Yamada, T. 2006. Acts of Command and Changing Obligations. In *Proceedings CLIMA VII*, eds. K. Inoue, K. Satoh and F. Toni. Also in Lecture Notes in AI, 4371 (2007), 1–19. Berlin: Springer.

Zarnic, B. 1999. Validity of Practical Inference. Research Report PP-1999-23, ILLC, University of Amsterdam.

Zvesper, J. 2008. What You Really Want. Working Paper, Gloriclass Center, ILLC, University of Amsterdam.

Chapter 4
Preference, Priorities and Belief

Dick de Jongh and Fenrong Liu

Abstract In this paper we consider preference over objects. We show how this preference can be derived from priorities, properties of these objects, a concept which is initially from optimality theory. We do this both in the case when an agent has complete information and in the case when an agent only has beliefs about the properties. After the single agent case we also consider the multi-agent case. In each of these cases, we construct preference logics, some of them extending the standard logic of belief. This leads to interesting connections between preference and beliefs. We strengthen the usual completeness results for logics of this kind to representation theorems. The representation theorems describe the reasoning that is valid for preference relations that have been obtained from priorities. In the multi-agent case, these representation theorems are strengthened to the special cases of cooperative and competitive agents. We study preference change with regard to changes of the priority sequence, and change of beliefs. We apply the dynamic epistemic logic approach, and in consequence reduction axioms are presented. We conclude with some possible directions for future work.

4.1 Motivation

The notion of preference occurs frequently in game theory, decision theory, and many other research areas. Typically, preference is used to draw comparison between two alternatives explicitly. Studying preference and its general properties has become a main logical concern after the pioneering seminar work by [Hal57] and [Wri63] witness [Jen67], [Cre71], [Tra85], [DW94], [Han01], [BRG07] etc., and more recently work on dynamics of preference, e.g. [Han95] and [BL07]. Let us single out immediately the two distinctive characteristics of the approach to preference we take in this paper.

D. de Jongh
Institute for Logic, Language and Computation (ILLC), University of Amsterdam,
The Netherlands
e-mail: d.h.j.dejongh@uva.nl

F. Liu (✉)
Department of Philosophy, School of Humanities and Social Sciences, Tsinghua University, China
e-mail: fenrong@tsinghua.edu.cn

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

- Most of the previous work has taken preference to be a primitive notion, without considering how it comes into being. We take a different angle here and explore both preference and its origin. We think that preference can often be rationally derived from a more basic source, which we will call a *priority base*. In this manner we have two levels: the priority base, and the preference derived from it. We hope this new perspective will shed light on the reasoning underlying preference, so that we are able to discuss *why* we prefer one thing over another. There are many ways to get preference from such a priority base, a good overview can be found in [CMLLM04].
- In real life we often encounter situations in which no complete information is available. Preference will then have to be based on our beliefs, i.e. do we believe certain properties from the priority base to apply or not? Apparently, this calls for a combination of doxastic language and preference language. We will show a close relationship between preference and beliefs. To us, both are mental attitudes. If we prefer something, we believe we do (and conversely). In addition, this paper is also concerned with the dynamics of preference. By means of our approach, we can study preference changes, whether they are due to a change in the priority base, or caused by belief revision.

Depending on the actual situation, preference can be employed to compare alternative states of affairs, objects, actions, means, and so on, as listed in [Wri63]. One requirement we impose is that we consider only mutually exclusive alternatives. In this paper, we consider in first instance preference over *objects* rather than between propositions (compare [DW94]). Objects are, of course, congenitally mutually exclusive. Although the priority base approach is particularly well suited to compare preference between objects, it can be applied to the study of the comparison of other types of alternatives as well. When comparing objects, the kind of situation to be thought of is:

Example 4.1.1. Alice is going to buy a house. For her there are several things to consider: the cost, the quality and the neighborhood, strictly in that order. All these are clear-cut for her, e.g. the cost is good if it is inside her budget, otherwise it is bad. Her decision is then determined by the information whether the alternatives have the desirable properties, and by the given order of importance of the properties.

In other words, Alice's preference regarding houses is derived from the priority order of the properties she considers. This paper aims to propose a logic to model such situations. When covering situations in which Alice's preference is based on incomplete information belief will enter into the logic as an operation.

There are several points to be stressed beforehand, in order to avoid misunderstandings: First, our intuition of priority base is linked to graded semantics, e.g. spheres semantics by [Lew73]. We take a rather syntactical approach in this paper, but that is largely a question of taste, one can go about it semantically as well. We will return to this point several times. Second, we will mostly consider a linearly ordered priority base. This is simple, giving us a quasi-linear order of preference. But our approach can be adapted to the partially ordered case, as we will indicate in Section 4.3. Third, when we add a belief operator to the preference language (fragment of *FOL*), it may seem that we are heading into doxastic predicate logic. This is

true, but we are not going to be affected by the existing difficult issues in that logic. What we are using in this context is a very limited part of the language. Finally, although we start with a two level perspective this results on the preference side in logics that are rather like ordinary propositional modal logics. The bridge between the two levels is then given by theorems that show that any models of these modal logics can be seen as having been constructed from a priority base. These theorems are a kind of completeness theorems, but we call them *representation theorems* to distinguish them from the purely modal completeness results.

The following sections are structured as follows: In Section 4.2, we start with a simple language to study the rigid case in which the priorities lead to a clear and unambiguous preference ordering. In Section 4.3 we review some basics about ordering. In Section 4.4, a proof of a representation theorem for the simple language without beliefs is presented. Section 4.5 considers what happens when the agent has incomplete information about the priorities with regard to the alternatives. In Section 4.6 we will look at changes in preference caused by two different sources: changes in beliefs, and changes of the sequence of priorities. Section 4.7 is an extension to the multi-agent system. We will prove representation theorems for the general case, and for the special cases of cooperative agents and competitive agents. Finally, we end up with a few conclusions and remarks about possible future work.

4.2 From Priorities to Preference

As we mentioned in the preceding, there are many ways to derive preference from the priority base. We choose one of the mechanisms, the way of Optimality Theory (*OT*, cf. [PS93]), as an illustration because we like the intuition behind this mechanism. Along the way, we will discuss other approaches to indicate how our method can be applied to them just as well.

Here is a brief review of some ideas from optimality theory that are relevant to the current context. In optimality theory a set of conditions is applied to the alternatives generated by the grammatical or phonological theory, to produce an optimal solution. It is by no means sure that the optimal solution satisfies all the conditions. There may be no such alternative. The conditions, called *constraints*, are strictly ordered according to their importance, and the alternative that satisfies the earlier conditions best (in a way described more precisely below) is considered to be the optimal one. This way of choosing the optimal alternative naturally induces a preference ordering among all the alternatives. We are interested in formally studying the way the constraints induce the *preference ordering* among the alternatives. The attitude in our investigations is somewhat differently directed than in optimality theory.[1]

[1] Note that in optimality theory the optimal alternative is chosen unconsciously; we are thinking mostly of applications where conscious choices are made. Also, in optimality theory the application of the constraints to the alternatives lead to a *clear* and *unambiguous* result: either the constraint

Back to the issues of preference, to discuss preference over objects, we use a first order logic with constants $d_0, d_1, \ldots$; variables $x_0, x_1, \ldots$; and predicates $P, Q, P_0, P_1, \ldots$. In practice, we are thinking of finite domains, monadic predicates, simple formulas, usually quantifier free or even variable free. The following definition is directly inspired by optimality theory, but to take a neutral stance we use the words priority sequence instead of constraint sequence.

Definition 4.2.1. A *priority sequence* is a finite ordered sequence of formulas (*priorities*) written as follows:

$$C_1 \gg C_2 \cdots \gg C_n \quad (n \in \mathbb{N}),$$

where each of C_m $(1 \leq m \leq n)$ is a formula from the language, and there is exactly one free variable x, which is a common one to each C_m.

We will use symbols like $\mathfrak{C}$ to denote priority sequences. The priority sequence is linearly ordered. It is to be read in such a way that the earlier priorities count strictly heavier than the later ones, e.g. $C_1 \wedge \neg C_2 \wedge \cdots \wedge \neg C_m$ is preferable over $\neg C_1 \wedge C_2 \wedge \cdots \wedge C_m$ and $C_1 \wedge C_2 \wedge C_3 \wedge \neg C_4 \wedge \neg C_5$ is preferable over $C_1 \wedge C_2 \wedge \neg C_3 \wedge C_4 \wedge C_5$. A difference with optimality theory is that we look at *satisfaction* of the priorities whereas in optimality theory *infractions* of the constraints are stressed. This is more a psychological than a formal difference. However, optimality theory knows multiple infractions of the constraints and then counts the number of these infractions. We do not obtain this with our simple objects, but we think that possibility can be achieved by considering composite objects, like strings.

Definition 4.2.2. Given a priority sequence of length n, two objects x and y, *Pref(x,y)* is defined as follows:

$Pref_1(x, y) ::= C_1(x) \wedge \neg C_1(y)$,
$Pref_{k+1}(x, y) ::= Pref_k(x, y) \vee (Eq_k(x, y) \wedge C_{k+1}(x) \wedge \neg C_{k+1}(y)), k < n$,
$Pref(x, y) ::= Pref_n(x, y)$,

where the auxiliary binary predicate $Eq_k(x, y)$ stands for $(C_1(x) \leftrightarrow C_1(y)) \wedge \cdots \wedge (C_k(x) \leftrightarrow C_k(y))$.[2]

In Example 4.1.1, Alice has the following priority sequence:

$$C(x) \gg Q(x) \gg N(x),$$

where $C(x)$, $Q(x)$ and $N(x)$ are intended to mean 'x has low cost', 'x is of good quality' and 'x has a nice neighborhood', respectively. Consider two houses

clearly is true of the alternative or it is not, and that is something that is not sensitive to change. We will loosen this condition and consider issues that arise when changes do occur.

[2] This way of deriving an ordering from a priority sequence is called *leximin ordering* in [CMLLM04].

d_1 and d_2 with the following properties: $C(d_1), C(d_2), \neg Q(d_1), \neg Q(d_2), N(d_1)$ and $\neg N(d_2)$. According to the above definition, Alice prefers d_1 over d_2, i.e. $Pref(d_1, d_2)$.

Unlike later, in Section 4.5, belief does not enter into this definition. This means that $Pref(x, y)$ can be read as *x is superior to y*, or *under complete information x is preferable over y*.

Remark 4.2.3. Our method easily applies when the priorities become graded. Take the Example 4.1.1, if Alice is more particular, she may split the cost C into C^1 very low cost, C^2 low cost, C^3 medium cost, similarly for the other priorities. The original priority sequence $C(x) \gg Q(x) \gg N(x)$ may change into

$$C^1(x) \gg C^2(x) \gg Q^1(x) \gg C^3(x) \gg Q^2(x) \gg N^1(x) \gg \dots .$$

As we mentioned at the beginning, we have chosen a syntactic approach expressing priorities by formulas. If we switch to a semantical point of view, the priority sequence translates into pointing out a sequence of n sets in the model. The elements of the model will be objects rather than worlds as is usual in this kind of study. But one should see this really as an insignificant difference. If one prefers, one may for instance in Example 4.1.1 replace house d by the situation in which Alice has bought the house d.

When one points out sets in a model, Lewis' sphere semantics ([Lew73] pp 98–99) comes to mind immediately. The n sets in the model obtained from the priority base are in principle unrelated. In the sphere semantics the sets which are pointed out are linearly ordered by inclusion. To compare with the priority base we switch to a syntactical variant of sphere semantics, a sequence of formulas $G_1, \dots, G_m$ such that $G_i(x)$ implies $G_j(x)$ if $i \leq j$. These formulas express the preferability in a more direct way, $G_1(x)$ is the most preferable, $G_m(x)$ the least. The two approaches are equivalent in the sense that they can be translated into each other.

Theorem 4.2.4. *A priority sequence $C_1 \gg C_2 \cdots \gg C_m$ gives rise to a G-sequence of length 2^m. In the other direction a priority sequence can be obtained from a G-sequence logarithmic in the length of the G-sequence.*

Proof. Let us just look at the case that $m = 3$. Assuming that we have the priority sequence $C_1 \gg C_2 \gg C_3$, the preference of objects is decided by where their properties occur in the following list:

$R_1 : C_1 \wedge C_2 \wedge C_3$
$R_2 : C_1 \wedge C_2 \wedge \neg C_3$
$R_3 : C_1 \wedge \neg C_2 \wedge C_3$
$R_4 : C_1 \wedge \neg C_2 \wedge \neg C_3$
$R_5 : \neg C_1 \wedge C_2 \wedge C_3$
$R_6 : \neg C_1 \wedge C_2 \wedge \neg C_3$
$R_7 : \neg C_1 \wedge \neg C_2 \wedge C_3$
$R_8 : \neg C_1 \wedge \neg C_2 \wedge \neg C_3$

The $G_i s$ are constructed as disjunctions of members of this list. In their most simple form, they can be stated as follows:

$G_1 : R_1$
$G_2 : R_1 \vee R_2$
$\vdots$
$G_8 : R_1 \vee \ R_2 \cdots \vee R_8$

On the other hand, given a G_i-sequence, we can define C_i as follows,

$C_1 = R_1 \vee R_2 \vee R_3 \vee R_4$
$C_2 = R_1 \vee R_2 \vee R_5 \vee R_6$
$C_3 = R_1 \vee R_3 \vee R_5 \vee R_7$

And again this can be simply read off from a picture of the G-spheres. The relationship between C_i, R_i, and G_i can be seen from the Fig. 4.1. ■

Remark 4.2.5. In applying our method to such spheres, the definition of $\underline{Pref}(x, y)$ comes out to be $\forall i\,(y \in G_i \rightarrow x \in G_i)$. The whole discussion implies of course that our method cannot only be applied to sphere models but also to any other approach which can be reduced to sphere models.

Remark 4.2.6. As we pointed out at the beginning, one can define preference from a priority sequence $\mathfrak{C}$ in various different ways, all of which we can handle. Here is one of these ways, called *best-out ordering* in [CMLLM04], as an illustration. We define the preference as follows:

$$Pref(x, y) \quad \text{iff} \quad \exists C_j \in \mathfrak{C}(\forall C_i \gg C_j\,((C_i(x) \wedge C_i(y)) \wedge (C_j(x) \wedge \neg C_j(y))).$$

Now we only continue along the priority sequence as long as we receive positive information. Returning the Example 4.1.1, this means that under this option we only get the conclusion that $\underline{Pref}(d_1, d_2)$ and $\underline{Pref}(d_2, d_1)$: d_1 and d_2 are equally preferable, because after observing that $\neg Q(d_1), \neg Q(d_2)$, Alice won't consider N at all.

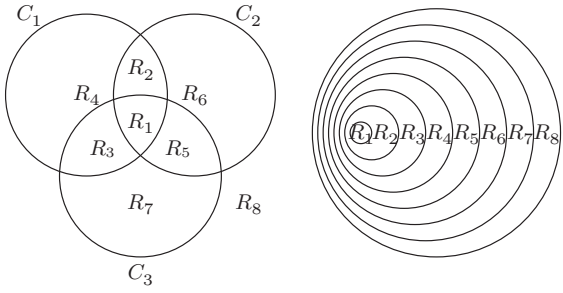

Fig. 4.1 C_i, R_i, and G_i

4.3 Order

In this section we will just run through the types of order that we will use. A relation $<$ is a *linear order* if $<$ is irreflexive, transitive and asymmetric ($x < y \rightarrow \neg(y < x)$), and satisfies *totality*:

$$x < y \vee x = y \vee y < x$$

More precisely, $<$ is called a *strict* linear order. A *non-strict* linear order $\leq$ is a reflexive, transitive, antisymmetric ($x \leq y \wedge y \leq x \rightarrow x = y$), and total relation. It is for various reasons useful to introduce both variants of orderings.

Mathematically, strict and non-strict linear orders can easily be translated into each other:

(1) $x < y \leftrightarrow x \leq y \wedge x \neq y$, or
(2) $x < y \leftrightarrow x \leq y \wedge \neg(y \leq x)$
(3) $x \leq y \leftrightarrow x < y \vee x = y$, or
(4) $x \leq y \leftrightarrow x < y \vee (\neg(x < y) \wedge \neg(y < x))$

Optimality theory only considers linearly ordered constraints. These will be seen to lead to a *quasi-linear order* of preferences, i.e. a relation $\preccurlyeq$ that satisfies all the requirements of a non-strict linear order but antisymmetry. A quasi-linear ordering contains *clusters* of elements that are 'equally large'. Such elements have the relation $\leq$ to each other. Most naturally one would take for the strict variant $\prec$ an irreflexive, transitive, total relation. If one does that, strict and non-strict orderings can still be translated into each other (only by using alternatives (2) and (4) above though, not (1) and (3)).

However, *Pref* is normally taken to be a strict order, i.e. an asymmetric relation, and we agree with that, so we take the option of $\prec$ as an irreflexive, transitive, asymmetric relation. Then $\prec$ is definable in terms of $\preccurlyeq$ by use of (2), but not $\preccurlyeq$ in terms of $\prec$. That is clear from the picture below, an irreflexive, transitive, asymmetric relation cannot distinguish between the two given orderings (Fig. 4.2).

One needs an additional equivalence relation $x \sim y$ to express that x and y are elements in the same cluster; $x \sim y$ can be defined by

(5) $x \sim y \leftrightarrow x \leq y \wedge y \leq x$

Then, in the other direction, $x \leq y$ can be defined in terms of $<$ and $\sim$:

(6) $x \leq y \leftrightarrow x < y \vee x \sim y$

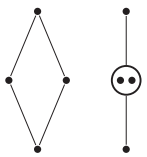

Fig. 4.2 Incomparability and indifference

It is certainly possible to extend our discussion to partially ordered sets of priorities. We will not really pursue this in this paper, but let us spend a few words on the issue. The preference relation will no longer be a quasi-linear order, but a so-called *quasi-order*: in the non-strict case a reflexive and transitive relation, in the strict case an asymmetric, transitive relation. One can still use (2) to obtain a strict quasi-order from a non-strict one and (6) to obtain a non-strict quasi-order from a strict one and $\sim$. However, we will see in Section 4.5 that in some contexts involving beliefs these translations no longer give the intended result. In such a case one has to be satisfied with the fact that (5) still holds and that $\prec$ as well as $\sim$ imply $\preccurlyeq$.

One will in practice meet partially ordered priority sequences when there are several priorities of incomparable strength. Take the Example 4.1.1 again, where now instead of just three properties to consider, Alice also takes the 'transportation convenience' into account. But for her neighborhood and transportation convenience are really incomparable. Abstractly speaking, this indeed means that the priority sequence is now partially ordered. We show in the following how to define preference based on such a partially ordered priority sequence. We consider a set of priorities $C_1, \ldots, C_n$ with the relation $\gg$ between them a partial order.

Definition 4.3.1. We define $Pref_n(x, y)$ by induction, where $\{n_1, ..., n_k\}$ is the set of immediate predecessors of n.

$$\underline{Pref_n(x,y)} ::= \underline{Pref_{n_1}}(x,y) \wedge ... \wedge \underline{Pref_{n_k}}(x,y) \wedge ((C_n(y) \rightarrow C_n(x)) \vee (Pref_{n_1}(x,y) \vee ... \vee Pref_{n_k}(x,y)))$$

where as always $Pref_m(x, y) \leftrightarrow \underline{Pref_m}(x, y) \wedge \neg \underline{Pref_m}(y, x)$.

This definition again has the inductive form we favor. Moreover, we regard finite partial orders as the most important, and restricted to those, the definition is equivalent to the one in [Gro91] and [ARS95]. This connection has been investigated in [Liu08] too. For more discussion on the relation between partially ordered priorities and G-spheres, see [Lew81], for the important special case that the set of priorities is completely unordered (which is also a partial order of course), we refer to [Kra81].

4.4 A Representation Theorem

In the following we will write *Pref* for the strict version of preference, $\underline{Pref}$ for the non-strict version, and let *Eq* correspond to $\sim$, expressing two elements are equivalent. Clearly, no matter what the priorities are, the non-strict preference relation has the following general properties:

(a) $\underline{Pref}(x, x)$
(b) $\underline{Pref}(x, y) \vee \underline{Pref}(y, x)$
(c) $\underline{Pref}(x, y) \wedge \underline{Pref}(y, z) \rightarrow \underline{Pref}(x, z)$

(a), (b) and (c) express reflexivity, totality and transitivity, respectively. Thus, $\underline{Pref}$ is a quasi-linear relation; it lacks antisymmetry.

Unsurprisingly, (a), (b) and (c) are a complete set of principles for preference. We will put this in the form of a representation theorem as we announced in the introduction. In this case it is a rather trivial matter, but it is worthwhile to execute it completely as an introduction to the later variants. We reduce the first order language for preference to its core:

Definition 4.4.1. Let Γ be a set of propositional variables, and D be a finite domain of objects, the *reduced language* of preference logic is defined in the following,

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \underline{Pref}(d_i, d_j),$$

where p, d_i respectively denote elements from Γ and D.

The reduced language contains the propositional calculus. From this point onwards we refer to the language with variables, quantifiers, predicates as the *extended language*. In the reduced language, we rewrite the axioms as follows:

(a) $\underline{Pref}(d_i, d_i)$
(b) $\underline{Pref}(d_i, d_j) \vee \underline{Pref}(d_j, d_i)$
(c) $\underline{Pref}(d_i, d_j) \wedge \underline{Pref}(d_j, d_k) \rightarrow \underline{Pref}(d_i, d_k)$

We call this axiom system **P**.

Theorem 4.4.2. (*representation theorem*). $\vdash_{\mathbf{P}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences.*

Proof. The direction from left to right is obvious. Assume formula $\varphi(d_1, \ldots, d_n, p_1, \ldots, p_k)$ is not derivable in **P**. Then a non-strict quasi-linear ordering of the $d_1, \ldots, d_n$ exists, which, together with a valuation of the atoms $p_1, \ldots, p_k$ in φ falsifies $\varphi(d_1, \ldots, d_n)$. Let us just assume that we have a linear order (adaptation to the more general case of quasi-linear order is simple), and also, w.l.o.g. that the ordering is $d_1 > d_2 > \cdots > d_n$. Then we introduce an extended language containing unary predicates $P_1, \ldots, P_n$ with a priority sequence $P_1 \gg P_2 \cdots \gg P_n$ and let P_i apply to d_i only. Clearly then the preference order of $d_1, \ldots, d_n$ with respect to the given priority sequence is from left to right. We have transformed the model into one in which the defined preference has the required properties.[3] ■

Remark 4.4.3. It is instructive to execute the above proof for the reduced language containing some additional predicates $Q_1, \ldots, Q_k$. One would like then to obtain a priority sequence of formulas in the language built up from Q_1 to Q_k. This is possible if in the model $\mathcal{M}$ each pair of constants d_i and d_j is distinguishable by formulas in this language, i.e. for each i and j $(i \neq j)$, there exists a formula φ_{ij} such that

[3] Note that, although we used n priorities in the proof to make the procedure easy to describe, in general $log_2(n) + 1$ priorities are sufficient for the purpose.

$\mathcal{M} \models \varphi_{ij}(d_i)$and $\mathcal{M} \models \neg\varphi_{ij}(d_j)$. In such a case, the formula $\psi_i = \bigwedge_{i \neq j} \varphi_{ij}$ satisfies only d_i. And $\psi_1 \gg \cdots \gg \psi_n$ is the priority sequence as required. It would be necessary to introduce new predicates when two constants are indistinguishable. A trivial method to do this is to allow identity in the language, $x = d_1$ obviously distinguishes d_1 and d_2.

Let us at this point stress once more what the content of a representation theorem is. It tells us that the way we have obtained the preference relations, namely from a priority sequence, does not affect the general reasoning about preference, its logic. The proof shows this in a strong way: if we have a model in which the preference relation behaves in a certain manner, then we can think of this preference as derived from a priority sequence without disturbing the model as it is.

4.5 Preference and Belief

In this section, we discuss the situation that arises when an agent has only incomplete information, but she likes to express her preference. The language will be extended with belief operators $B\varphi$ to deal with such uncertainty, and it is a small fragment of doxastic predicate logic. It would be interesting to consider what more the full language can bring us, but we will leave this question to other occasions. We will take the standard **KD45** as the logic for beliefs (cf. [MvdH95]), though we are aware of the philosophical discussions on beliefs and the options of proper logical systems.

Interestingly, the different definitions of preference we propose in the following spell out different "procedures" an agent may follow to decide her preference when processing the incomplete information about the relevant properties. Which procedure is taken strongly depends on the domain or the type of agents. In the new language, the definition of priority sequence remains the same, i.e. a priority C_i is a formula from the language *without* belief operators.

Definition 4.5.1. (decisive preference). Given a priority sequence of length n, and two objects x and y, *Pref*(*x,y*) is defined as follows:

$$Pref_1(x, y) ::= BC_1(x) \wedge \neg BC_1(y),$$
$$Pref_{k+1}(x, y) ::= Pref_k(x, y) \vee (Eq_k(x, y) \wedge BC_{k+1}(x) \wedge \neg BC_{k+1}(y)), k < n,$$
$$Pref(x, y) ::= Pref_n(x, y),$$

where $Eq_k(x, y)$ stands for $(BC_1(x) \leftrightarrow BC_1(y)) \wedge \cdots \wedge (BC_k(x) \leftrightarrow BC_k(y))$.

To determine the preference relation, one just runs through the sequence of relevant properties to check whether one believes them of the objects. But at least two other options of defining preference seem reasonable as well.

Definition 4.5.2. (conservative preference). Given a priority sequence of length n, two objects x and y, *Pref(x,y)* is defined below:

$Pref_1(x, y) ::= BC_1(x) \wedge B\neg C_1(y)$,
$Pref_{k+1}(x, y) ::= Pref_k(x, y) \vee (Eq_k(x, y) \wedge BC_{k+1}(x) \wedge B\neg C_{k+1}(y)), k < n$,
$Pref(x, y) ::= Pref_n(x, y)$

where $Eq_k(x, y)$ stands for $(BC_1(x) \leftrightarrow BC_1(y)) \wedge (B\neg C_1(x) \leftrightarrow B\neg C_1(y)) \wedge \cdots \wedge (BC_k(x) \leftrightarrow BC_k(y)) \wedge (B\neg C_k(x) \leftrightarrow B\neg C_k(y))$.

Definition 4.5.3. (deliberate preference). Given a priority sequence of length n, two objects x and y, *Pref(x,y)* is defined below:

$Supe_1(x, y)$[4] $::= C_1(x) \wedge \neg C_1(y)$
$Supe_{k+1}(x, y) ::= Supe_k(x, y) \vee (Eq_k(x, y) \wedge C_{k+1}(x) \wedge \neg C_{k+1}(y)), k < n$
$Supe(x, y) ::= Supe_n(x, y)$
$Pref(x, y) ::= B(Supe(x, y))$

where $Eq_k(x, y)$ stands for $(C_1(x) \leftrightarrow C_1(y)) \wedge \cdots \wedge (C_k(x) \leftrightarrow C_k(y))$.

To better understand the difference between the above three definitions, we look at the Example 4.1.1 again, but in three different variations:

A. Alice favors Definition 4.5.1: She looks at what information she can get, she reads that d_1 has low cost, about d_2 there is no information. This immediately makes her decide for d_1. This will remains so, no matter what she hears about quality or neighborhood.
B. Bob favors Definition 4.5.2: The same thing happens to him. But he reacts differently than Alice. He has no preference, and that will remain so as long as he hears nothing about the cost of d_2, no matter what he hears about quality or neighborhood.
C. Cora favors Definition 4.5.3: She also has the same information. On that basis Cora cannot decide either. But some more information about quality and neighborhood helps her to decide. For instance, suppose she hears that d_1 has good quality or is in a good neighborhood, and d_2 is not of good quality and not in a good neighborhood. Then Cora believes that, no matter what, d_1 is superior, so d_1 is her preference. Note that such kind of information could not help Bob to decide.

Speaking more generally in terms of the behaviors of the above agents, it seems that Alice always decides what she prefers on the basis of the limited information she has. In contrast, Bob chooses to wait and require more information. Cora behaves somewhat differently, she first tries to do some reasoning with all the available information before making her decision. This suggests yet another perspective on diversity of agents than discussed in [BL04] and [Liu09].

Apparently, we have the following fact.

[4] Superiority is just defined as preference was in the previous section.

Fact 4.5.4

- *Totality holds for Definition 4.5.1, but not for Definitions 4.5.2 or 4.5.3;*
- *Among the above three definitions, Definition 4.5.2 is the strongest in the sense that if $Pref(x, y)$ holds according to Definition 4.5.2, then $Pref(x, y)$ holds according to Definitions 4.5.1 and 4.5.3 as well.*

It is striking that, if in Definition 4.5.3, one plausibly also defines $\underline{Pref}(x, y)$ as $B(\underline{Supe}(x, y))$, then the normal relation between $Pref$ and $\underline{Pref}$ no longer holds: $Pref$ is not definable in terms of $\underline{Pref}$, or even $\underline{Pref}$ in terms of $Pref$ and Eq.

For all three definitions, we have the following theorem.

Theorem 4.5.5. $\underline{Pref}(x, y) \leftrightarrow B\underline{Pref}(x, y)$.

Proof. In fact we prove something more general in **KD45**. Namely, if α is a propositional combination of B-statements, then $\vdash_{\mathbf{KD45}} \alpha \leftrightarrow B\alpha$. Since $\underline{Pref}(x, y)$ is in all three cases indeed a propositional combination of B-statements, and since we assume the principles of **KD45** to hold, this is sufficient.

From left to right, since α is a propositional combination of B-statements, it can be transformed into conjunctive normal form: $\beta_1 \vee \cdots \vee \beta_k$. It is clear that $\vdash_{\mathbf{KD45}} \beta_i \rightarrow B\beta_i$ for each i, because each member γ of the conjunction β_i implies $B\gamma$. If $A = \beta_1 \vee \cdots \vee \beta_k$ holds then some β_i holds, so $B\beta_i$, so $B\alpha$. Then we immediately have: $\vdash_{\mathbf{KD45}} \neg\alpha \rightarrow B\neg\alpha$ (*) as well, since $\neg\alpha$ is also a propositional combination of B-statements if α is.

From right to left: Suppose $B\alpha$ and $\neg\alpha$. Then $B\neg\alpha$ by (*), so $B\bot$, but this is impossible in **KD45**, therefore α holds. ■

Corollary 4.5.6. $\neg\underline{Pref}(x, y) \leftrightarrow B\neg\underline{Pref}(x, y)$.

Actually, we think it is proper that Theorem 4.5.5 and Corollary 4.5.6 hold because we believe that preference describes a state of mind in the same way that belief does. Just as one believes what one believes, one believes what one prefers.

We can generalize the representation result (Theorem 4.4.2) if we stick to Definition 4.5.1 (decisive preference). This definition is most congenial to us in any case. Let us consider the reduced language built up from standard propositional letters plus $\underline{Pref}(d_i, d_j)$, by the connectives and belief operators B. Again we have the normal principles of **KD45** for B.

Definition 4.5.7. The **KD45-P** system includes the principles below, plus *Modus ponens*(MP), as well as *Generalization* for the operator B.

(a) $\underline{Pref}(d_i, d_j)$
(b) $\underline{Pref}(d_i, d_j) \vee \underline{Pref}(d_j, d_i)$
(c) $\underline{Pref}(d_i, d_j) \wedge \underline{Pref}(d_j, d_k) \rightarrow \underline{Pref}(d_i, d_k)$
(1.) $\neg B\bot$
(2.) $B\varphi \rightarrow BB\varphi$
(3.) $\neg B\varphi \rightarrow B\neg B\varphi$
(4.) $\underline{Pref}(d_i, d_j) \leftrightarrow B\underline{Pref}(d_i, d_j)$

Definition 4.5.8. A model of **KD45-P** is a tuple $\langle W, D, R, \{\preceq_w\}_{w\in W}, V\rangle$, where W is a set of worlds, D is a set of constants, R is a euclidean and serial accessibility relation on W. Namely, it satisfies $\forall xyz((Rxy \wedge Rxz) \rightarrow Ryz)$ and $\forall x\exists yRxy$. For each w, $\preceq_w$ is a quasi-linear order on D, which is constant throughout each euclidean class, i.e., if wRw', then $a \preceq_w b$ iff $a \preceq_{w'} b$. V is an evaluation function in the ordinary manner.

We remind the reader that the set of worlds in a **KD45**-model is partitioned into what we will call *euclidean classes*. In most respects euclidean classes are like equivalence classes, but a number of points may be irreflexive and then have R relations just towards all the reflexive members (the *equivalence part*) of the class. The equivalence part is an equivalence class in the ordinary sense. It is also easy to see that, if w is a world in such a model, then the euclidean class in which w resides is the set $\{w'' \mid \exists w'(w''Rw' \wedge wRw')\}$. The reader can easily check that the principles of **KD45-P** are valid in the **KD45-P**-models.

Theorem 4.5.9. *The* **KD45-P** *system is complete.*

Proof. The canonical model of this logic **KD45-P** has the required properties given in Definition 4.5.8: The belief accessibility relation R is euclidean and serial. This means that with regard to R the model falls apart into euclidean classes. In each node $\underline{Pref}$ is a quasi-linear order of the constants. Note that, for totality, we rely on the fact that we are using Definition 4.5.1. Within a euclidean class the preference order is constant, by $B\underline{Pref} \leftrightarrow \underline{Pref}$. This suffices to prove completeness. ■

Theorem 4.5.10. *The logic* **KD45-P** *has the finite model property.*

Proof. By standard methods. ■

Theorem 4.5.11. (*representation theorem*). $\vdash_{\mathbf{KD45-P}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences.*

Proof. Suppose that $\nvdash_{\mathbf{KD45-P}} \varphi(d_1, ..., d_n, p_1, ..., p_m)$. Theorem 4.5.9, there is a model with a world w in which φ is falsified. We restrict the model to the euclidean class where w resides. (Note that, by the remarks above, this is a generated submodel.) Since the ordering of the constants is the same throughout euclidean classes, the ordering of the constants is now the same throughout the whole model. We can proceed as in Theorem 4.4.2 defining the predicates $P_1, \ldots, P_n$ in a constant manner throughout the model. Since we have a generated submodel, φ is still falsified in w. ■

Remark 4.5.12. The three definitions above are not the only definitions that might be considered. For instance, we can give a variation $(*)$ of Definition 4.5.2. For simplicity, we just use one predicate C.

$$Pref(x, y) ::= \neg B\neg C(x) \wedge B\neg C(y). \quad (*)$$

This means the agent can decide on her preference in a situation in which on the one hand she is not totally ready to believe $C(x)$, but considers it consistent with

what she assumes, on the other hand, she distinctly believes $\neg C(y)$. Compared with Definition 4.5.2, $(*)$ is weaker in the sense that it does not require explicit positive beliefs concerning $C(x)$.

We can even combine Definition 4.5.1 and $(*)$, obtaining the following:

$$Pref(x, y) ::= (BC(x) \wedge \neg BC(x)) \vee (\neg B\neg C(x) \wedge B\neg C(y). \quad (**)$$

Contrary to $(*)$, this gives a quasi-linear order.

4.6 Preference Changes

So far we have given different definitions for preference in a stable situation. Now we direct ourselves to changes in this situation. In the definition of preference in the presence of complete information, the only item subject to change is the priority sequence. In the case of incomplete information, not only the priority sequence, but also our beliefs can change. Both changes in priority sequence and changes in belief can cause preference change. In this section we study both. Note that priority change leads to a preference change in a way similar to entrenchment change in belief revision theory (see [Rot03]), but we take the methodology of dynamic epistemic logic in this context.

4.6.1 Preference Change Due to Priority Change

Let us first look at a variation of Example 4.1.1:

Example 4.6.1. Alice won a lottery prize of ten million dollars. Her situation has changed dramatically. Now she considers the quality most important.

In other words, the ordering of the priorities has changed. We will focus on the priority changes, and the preference changes they cause. To this purpose, we start by making the priority sequence explicit in the preference. We do this first for the case of complete information in language without belief. Let $\mathfrak{C}$ be a priority sequence with length n as in Definition 4.2.1. Then we write $Pref_{\mathfrak{C}}(x, y)$ for the preference defined from that priority sequence. Let us write $\mathfrak{C}^\frown C$ for adding C to the right of $\mathfrak{C}$, $C^\frown \mathfrak{C}$ for adding C to the left of $\mathfrak{C}$, $\mathfrak{C}^-$ for the sequence $\mathfrak{C}$ with its final element deleted, and finally, $\mathfrak{C}^{i \leftrightarrows i+1}$ for the sequence $\mathfrak{C}$ with its ith and i+1-th priorities switched. It is then clear that we have the following relationships:

$$Pref_{\mathfrak{C}^\frown C}(x, y) \leftrightarrow Pref_{\mathfrak{C}}(x, y) \vee (Eq_{\mathfrak{C}}(x, y) \wedge C(x) \wedge \neg C(y)),$$
$$Pref_{C^\frown \mathfrak{C}}(x, y) \leftrightarrow (C(x) \wedge \neg C(y)) \vee ((C(x) \leftrightarrow C(y)) \wedge Pref_{\mathfrak{C}}(x, y)),$$
$$Pref_{\mathfrak{C}^-}(x, y) \leftrightarrow Pref_{\mathfrak{C},n-1}(x, y),$$

$Pref_{\mathfrak{C}^{i \leftrightarrows i+1}}(x,y) \leftrightarrow Pref_{\mathfrak{C},i-1}(x,y) \vee (Eq_{\mathfrak{C},i-1}(x,y) \wedge C_{i+1}(x) \wedge \neg C_{i+1}(y)) \vee (Eq_{\mathfrak{C},i-1}(x,y) \wedge (C_{i+1}(x) \leftrightarrow C_{i+1}(y)) \wedge C_i(x) \wedge \neg C_i(y)) \vee (Eq_{\mathfrak{C},i+1}(x,y) \wedge Pref_{\mathfrak{C}}(x,y))$.

These relationships enable us to describe preference change due to changes of the priority sequence in the manner of dynamic epistemic logic (DEL). In DEL, the relationships between epistemic states under consideration before and after a change are represented by operators. These operators convert the state into its new form. Typically, the new state can be given completely in terms of the old state. This is captured by so called *reduction axioms*. We consider the operations $[^+C]$ of adding C to the right, $[C^+]$ of adding C to the left, $[-]$ of dropping the last element of a priority sequence of length n, $[i \leftrightarrow i{+}1]$ of interchanging the ith and i+1-th elements. Then we have the following reduction axioms:

$[^+C]Pref(x,y) \leftrightarrow Pref(x,y) \vee (Eq(x,y) \wedge C(x) \wedge \neg C(y))$,
$[C^+]Pref(x,y) \leftrightarrow ((C(x) \wedge \neg C(y)) \vee ((C(x) \leftrightarrow C(y)) \wedge Pref(x,y)))$,
$[-]Pref(x,y) \leftrightarrow Pref_{n-1}(x,y)$,
$[i \leftrightarrow i+1]Pref(x,y) \leftrightarrow Pref_{i-1}(x,y) \vee (Eq_{i-1}(x,y) \wedge C_{i+1}(x) \wedge \neg C_{i+1}(y)) \vee (Pref_i(x,y) \wedge (C_{i+1}(x) \leftrightarrow C_{i+1}(y))) \vee (Eq_{i+1}(x,y) \wedge Pref(x,y))$.

Of course, the first two are the more satisfactory ones, as the right hand side is constructed solely on the basis of the previous $Pref$ and the added priority C. Note that one of the first two, plus the third and the fourth are sufficient to represent any change whatsoever in the priority sequence. Noteworthy is that operator $[C^+]$ has exactly the same effects on a model as the operator $[\sharp C]$ in [BL07].

In the context of incomplete information when we have the language of belief, we can obtain similar reduction axioms for Definitions 4.5.1 and 4.5.2. For instance, for Definition 4.5.1, we need only replace C by BC and $\neg C$ by $\neg BC$. For Definition 4.5.3, the situation is very complicated, reduction axioms are simply not possible. To see this, we return to the Example of Cora. Suppose Cora has a preference on the basis of cost and quality, and she also has the given information relating quality and neighborhood. Then her new preference after 'neighborhood' has been adjoined to the priority sequence is not a function of her previous preference and her beliefs about the neighborhood. The beliefs relating quality and neighborhood are central for her reasoning, but they are neither contained in the beliefs supporting her previous preference, nor in the beliefs about the neighborhood per se.

4.6.2 Preference Change Due to Belief Change

Now we move to the other source which causes preference change, namely, a change in belief. Such a thing often occurs in real life, new information comes in, one changes one's beliefs. Technically, the update mechanisms of [BS06] and [Ben07] can immediately be applied to our system with belief. As preference is defined in terms of beliefs, we can calculate preference changes from belief change. We distinguish the two cases that the belief change is caused by an update with so-called *hard* information and that it is caused by an update with *soft* information.

4.6.2.1 Preference Change Under Hard Information

Consider a simpler version of the Example 4.1.1:

Example 4.6.2. This time Alice only considers the houses' cost (C) and their neighborhood (N) with $C(x) \gg N(x)$. There are two houses d_1 and d_2 available. The real situation is that $C(d_1), N(d_1), C(d_2)$ and $\neg N(d_2)$. First Alice prefers d_2 over d_1 because she believes $C(d_2)$ and $N(d_1)$. However, now Alice reads that $C(d_1)$ in a newspaper. She accepts this information, and accordingly changes her preference.

Here we assume that Alice treats the information obtained as hard information. She simply adds new information to her stock of beliefs. Figure 4.3 shows the situation before Alice's reading.

The figure can be read as a **KD45**-model. As usual, the dotted line denotes that Alice is uncertain about the two situations. In particular, she does not know whether $C(d_1)$ holds or not. After she reads that $C(d_1)$, the situation becomes Fig. 4.4. The $\neg C(d_1)$-world is eliminated from the model: Alice has updated her beliefs. Now she prefers d_1 over d_2.

We have assumed that we are using the elimination semantics (e.g. [Ben06]; [FHMV95], etc.) in which public announcement of the sentence A leads to the elimination of the $\neg A$ worlds from the model. We have the reduction axiom:

$$[!A]Pref_{\mathfrak{C}}(x, y) \leftrightarrow A \rightarrow Pref_{A \rightarrow \mathfrak{C}}(x, y),$$

where, if $\mathfrak{C}$ is the priority sequence $C_1 \gg \cdots \gg C_n$, $A \rightarrow \mathfrak{C}$ is defined as $A \rightarrow C_1 \gg \cdots \gg A \rightarrow C_n$.

We can go even further if we use conditional beliefs $B^{\psi}\varphi$ as introduced in [Ben07], with the meaning that φ is believed under the condition ψ. This immediately leads to the opportunity to introduce *conditional preference* $Pref^{\psi}(x, y)$ as well, by replacing B in the definitions in Section 4.5 by B^{ψ}. Assuming A is a formula without belief operators, an easy calculation gives us another form of the reduction axiom:

$$[!A]Pref(x, y) \leftrightarrow A \rightarrow Pref^{A}(x, y).$$

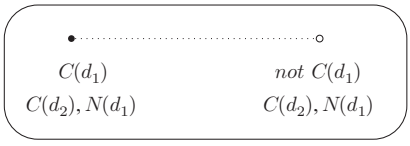

Fig. 4.3 Initial model

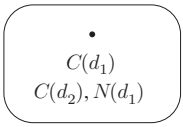

Fig. 4.4 Updated model

4.6.2.2 Preference Change Under Soft Information

When one meets information that is less solid, one needs a more subtle reaction to the information than simply adding it to one's stock of beliefs. One tends to believe the incoming information without discounting the possibility that it might be false. Let us switch to a semantical point of view for a moment. To discuss the impact of such so-called soft information on beliefs, the models are graded by a plausibility ordering $\leq$. For the one agent case one may just as well consider the model to consist of one euclidean class. The ordering of this euclidean class is such that the worlds in the equivalence part are the most plausible worlds. This means that for all the worlds w in the equivalence part and all the worlds u outside it, $w < u$. Otherwise $v < v'$ can only obtain between worlds outside the equivalence part. To be able to refer to the elements in the model, instead of only to the worlds accessible by the R-relation, we introduce the universal modality U and its dual E.

For the update by soft information, there are various nonequivalent approaches available, we choose the *lexicographic upgrade* $\Uparrow A$ introduced by [Vel96] and [Rot06], adopted by [Ben07] for this purpose. After the incoming information A, the $\neg A$-worlds are not deleted as in the case of hard information, one just updates the ordering $\leq$ by making all A-worlds strictly better than all $\neg A$-worlds, keeping among the A-worlds the old orders intact and doing the same for the $\neg A$-worlds. After the update the R-relations just point to the best A-worlds. The reduction axiom for belief proposed on this basis in [Ben07] is:

$$[\Uparrow A]B\varphi \leftrightarrow (EA \wedge B^A([\Uparrow A]\varphi)) \vee (\neg EA \wedge B([\Uparrow A]\varphi)).$$

Applying this to formulas φ which do not contain belief operators, one obtains for this restricted case a simpler form:

$$[\Uparrow A]B\varphi \leftrightarrow (EA \wedge B^A\varphi) \vee (\neg EA \wedge B\varphi).$$

Realizing that preference formulas are propositional combinations of this simple form one easily derives the reduction axiom for preference:

$$[\Uparrow A]Pref(x, y) \leftrightarrow (EA \wedge Pref^A(x, y)) \vee (\neg EA \wedge Pref(x, y)).$$

Or in a form closer to the one for hard information:

$$[\Uparrow A]Pref(x, y) \leftrightarrow (EA \rightarrow Pref^A(x, y)) \wedge (\neg EA \rightarrow Pref(x, y)).$$

The reduction axiom for conditional preference is:

$$[\Uparrow A]Pref^{\psi}(x, y) \leftrightarrow (E(A \wedge \psi) \rightarrow Pref^{A\wedge\psi}(x, y)) \wedge (\neg E(A \wedge \psi) \rightarrow Pref^{\psi}(x, y)).$$

As always in dynamic epistemic/doxastic logic the fact that we now have reduction axioms here implies that the completeness result in [Ben07] for dynamic belief logic can be extended to a dynamic preference logic. We will not spell out the details here.

4.7 Extension to the Many Agent Case

This section extends the results of Section 4.5 to the many agent case. This will generally turn out to be more or less a routine matter. But at the end of the section, we will see that the priority base approach gives us a start of an analysis of cooperation and competition of agents. We consider agents here as cooperative if they have the same goals (priorities), competitive if they have opposite goals. This is of course rather rudimentary because there are no actions in our models, but an important matter will be noticed immediately. That two agents have the same priority sequence does in no way imply that they agree on everything. Take for example two party members who agree exactly on the qualifications the candidate of their party should have (priorities). Still, they may not agree at all on how (they believe) a particular candidate satisfies these qualifications. Or, if Alice and her husband Bob are in perfect union about the requirements their new house should satisfy, still they may have a vehement disagreement whether a particular house satisfies these requirements: Alice may believe it is of good quality, but Bob doesn't. Even in this rudimentary approach the complexities of cooperation become clear. The way we define the concept of opposite goals for competitive agents (see just before Theorem 4.7.9) foreshadows the direction one may take to apply our approach to games. The language we are using is defined as follows.

Definition 4.7.1. Let Γ be a set of propositional variables, G be a group of agents, and D be a finite domain of objects, the *reduced language* of preference logic for many agents is defined in the following,

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \underline{Pref^a}(d_i, d_j) \mid B^a\varphi$$

where p, a, d_i respectively denote elements from Γ, G, and D.

Similarly to $\underline{Pref^a}$ expressing non-strict preference, we will use $Pref^a$ to denote the strict version. When we want to use the extended language, we add variables and the statements $P(d_i)$.

Definition 4.7.2. A *priority sequence* for an agent a is a finite ordered sequence of formulas written as follows: $C_1 \gg_a C_2 \cdots \gg_a C_n \quad (n \in \mathbb{N})$, where each C_m $(1 \leq m \leq n)$ is a formula using the predicates of the extended language of Definition 4.7.1, with one single free variable x, but without $\underline{Pref}$ and B.

Here we take decisive preference to define an agent's preference. But the results of this section apply to other definitions as well. It seems quite reasonable to allow in

this definition of $Pref^a$ formulas in the priority sequence that contain B^b and $Pref^b$ for agents b other than a. But we leave this for a future occasion.

Definition 4.7.3. Given a priority sequence of length n, two objects x and y, $Pref^a(x, y)$ is defined as follows:

$Pref^a_1(x, y) ::= B^a C_1(x) \wedge \neg B^a C_1(y)$,
$Pref^a_{k+1}(x, y) ::= Pref^a_k(x, y) \vee (Eq_k(x, y) \wedge B^a C_{k+1}(x) \wedge \neg B^a C_{k+1}(y)), k < n$,
$Pref^a(x, y) ::= Pref^a_n(x, y)$,

where $Eq_k(x, y)$ stands for $(B^a C_1(x) \leftrightarrow B^a C_1(y)) \wedge \cdots \wedge (B^a C_k(x) \leftrightarrow B^a C_k(y))$.

Definition 4.7.4. The preference logic for many agents **KD45-P$^{\mathbf{G}}$** is defined as follows,

(a) $\underline{Pref^a}(d_i, d_i)$
(b) $\underline{Pref^a}(d_i, d_j) \vee \underline{Pref^a}(d_j, d_i)$
(c) $\underline{Pref^a}(d_i, d_j) \wedge \underline{Pref^a}(d_j, d_k) \rightarrow \underline{Pref^a}(d_i, d_k)$
(1.) $\neg B^a \bot$
(2.) $B^a \varphi \rightarrow B^a B^a \varphi$
(3.) $\neg B^a \varphi \rightarrow B^a \neg B^a \varphi$
(4.) $\underline{Pref^a}(d_i, d_j) \leftrightarrow B^a \underline{Pref^a}(d_i, d_j)$

As usual, it also includes *Modus ponens*(MP), as well as *Generalization* for the operator B^a. It is easy to see that the above principles are valid for $\underline{Pref^a}$ extracted from a priority sequence.

Theorem 4.7.5. *The preference logic for many agents* **KD45-P$^{\mathbf{G}}$** *is complete.*

Proof. The canonical model of this logic **KD45-P$^{\mathbf{G}}$** has the required properties: The belief accessibility relations R_a are euclidean and serial. This means that with regard to R_a the model falls apart into a-euclidean classes. Again, in each node $Pref^a$ is a quasi-linear order of the constants and within an a-euclidean class the a-preference order is constant. This quasi-linearity and constance are of course the required properties for the preference relation. Same for the other agents. This shows completeness of the logic. ■

Theorem 4.7.6. *The logic* **KD45-P$^{\mathbf{G}}$** *has the finite model property.*

Proof. By standard methods. ■

A representation theorem can be obtained by showing that the model could have been obtained from priority sequences $C_1 \gg_a C_2 \cdots \gg_a C_m (m \in \mathbb{N})$ for all the agents.

Theorem 4.7.7. (*representation theorem*). $\vdash_{\mathbf{KD45-P^G}} \varphi$ *iff* φ *is valid in all models with each* $\underline{Pref^a}$ *obtained from a priority sequence.*

Proof. Let there be k agents $a_0, \ldots, a_{k-1}$. We provide each agent a_j with her own priority sequence $P_{n\times j+1} \gg_{a_j} P_{n\times j+2} \gg_{a_j} \ldots \gg_{a_j} P_{n\times(j+1)}$. From the previous proofs of representation theorems it is clear that it is sufficient to show that any model for **KD45-P$^{\mathbf{G}}$** for the reduced language can be extended by valuations for the $P_j(d_i)$'s in such a way that the preference relations are preserved. For each a_j-euclidean class, we follow the same procedure for $d_1, \ldots, d_n$ w.r.t. $P_{n\times j+1}, P_{n\times j+2}, \ldots, P_{n\times(j+1)}$ as in Theorem 4.4.2 w.r.t. $P_1, \ldots, P_n$. The preference orders obtained in this manner are exactly the $Pref^{a_j}$ relations in the model. ■

In the above case, the priority sequences for different agents are separate, and thus very different. Still stronger representation theorems can be obtained by requiring that the priority sequences for different agents are related, e.g. in the case of *cooperative agents*, that they are equal. We will consider the two agent case in the following.

Theorem 4.7.8. (*for two cooperative agents*). $\vdash_{\mathbf{KD45-P^G}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences shared by two cooperative agents.*

Proof. The two agents are a and b. We now have the priority sequence $P_1 \gg_a P_2 \gg_a \ldots \gg_a P_n$, same for b. It is sufficient to show that any model $\mathcal{M}$ with worlds W for **KD45-P$^{\mathbf{G}}$** for the reduced language can be extended by valuations for the $P_j(d_i)$'s in such a way that the preference relations are preserved. But, it is clear that in this case we cannot hope to do this purely on the model as it is because then from their shared priority sequence a and b would get the same preferences. We will get around this difficulty by enlarging the model, and obtaining what we want on the original part.

We start by making all $P_j(d_i)$'s true everywhere in the model. Next we extend the model as follows. For each a-euclidean class E in the model carry out the following procedure. Extend $\mathcal{M}$ with a complete isomorphic copy $\mathcal{M}_E = \{v_E \mid v \in W\}$ of $\mathcal{M}$ for all of the reduced language i.e. without the predicates P_j. Add R_a relations from any of the w in E to the copies v_E such that $w\, R_a\, v$. Now carry out the same procedure as in the proof of Theorem 4.4.2, just in E's copy E_E. What we do with regard to the P's in the rest of $\mathcal{M}_E$ is completely irrelevant. Now, in any w in $\mathcal{M}$, a will believe in $P_j(d_i)$ exactly as in the model in the proof of Theorem 4.7.7: the overall truth of the $P_j(d_i)$ in the a-euclidean class E in the original model has been made irrelevant. Thus, the preference orders obtained in this manner are exactly the $Pref^a$ relations in the model.

Next, do the same thing for b: add for each b-euclidean class F in $\mathcal{M}$ a whole new copy $\mathcal{M}_F$, and repeat the procedure followed for a. Both a and b will have preferences with regard to the same priority sequence. (But as noted before these preferences may be quite different.)

Finally, one notes that all formulas in the reduced language keep their original valuation on w in $\mathcal{M}$, because the model $\mathcal{M}$ is bisimilar for the reduced language to the new model consisting of $\mathcal{M}$ plus all the $\mathcal{M}_E$ and $\mathcal{M}_F$. The bisimulation simply consists of all pairs (v, w) where $w = v$, or $w = v_E$ or $w = v_F$ for some E or F. ■

For *competitive agents* we assume that if agent a has a priority sequence $D_1 \gg_a D_2 \gg \cdots \gg_a D_m (m \in \mathbb{N})$, then the opponent b has priority sequence $\neg D_m \gg_b \neg D_{m-1} \gg \cdots \gg_b \neg D_1$. These two priority sequences are such that under complete information they will order a set of objects in exactly the opposite manner.

Theorem 4.7.9. (*for two competitive agents*). $\vdash_{\mathbf{KD45-P^G}} \varphi$ *iff* φ *is valid in all models obtained from priority sequences for competitive agents.*

Proof. Let's assume two agents a and b. For a we take a priority sequence $P_1 \gg_a P_2 \gg_a \cdots \gg_a P_n \gg_a P_{n+1} \gg_a \cdots \gg_a P_{2n}$, and for b, we take $\neg P_{2n} \gg_b \neg P_{2n-1} \gg_b \cdots \gg_b \neg P_n \gg_b \neg P_{n-1} \gg_b \cdots \gg_b \neg P_1$. It is sufficient to show that any model $\mathcal{M}$ with worlds W for **KD45-P$^{\mathbf{G}}$** for the reduced language can be extended by valuations for the $P_j(d_i)$'s in such a way that the preference relations are preserved. We start by making all $P_1(d_i) \ldots P_n(d_i)$ true everywhere in the model and $P_{n+1}(d_i) \ldots P_{2n}(d_i)$ all false everywhere in the model. Next we extend the model as follows.

For each a-euclidean class E in the model carry out the following procedure. Extend $\mathcal{M}$ with a complete copy $\mathcal{M}_E$ of $\mathcal{M}$ for all of the reduced language i.e. without the predicates P_j. Add R_a relations from any of the w in E to the copies v_E such that $w R_a v$. Now define the values of the $P_1(d_i) \ldots P_n(d_i)$ in E_E as in the previous proof and make all $P_m(d_i)$ true everywhere for $m > n$. The preference orders obtained in this manner are exactly the $Pref^a$ relations in the model.

For each b-euclidean class F in the model carry out the following procedure. Extend $\mathcal{M}$ with a complete copy $\mathcal{M}_F$ of $\mathcal{M}$ for all of the reduced language i.e. without the predicates P_j. Add R_b relations from any of the w in F to the copies v_F such that $w R_b v$. Now define the values of the $\neg P_{2n}(d_i) \ldots \neg P_{n+1}(d_i)$ in F_F as for $P_1(d_i) \ldots P_n(d_i)$ in the previous proof and make all $P_m(d_i)$ true everywhere for $m \leq n$. The preference orders obtained in this manner are exactly the $Pref^b$ relations in the model.

Finally, one notes that all formulas in the reduced language keep their original valuation on w in $\mathcal{M}$, because the model $\mathcal{M}$ is bisimilar for the reduced language to the new model consisting of $\mathcal{M}$ plus all the M_E and M_F. ■

Remark 4.7.10. These last representation theorems are both a sign of strength and a sign of weakness of our systems. The weakness here is that they show that cooperation and competition cannot be differentiated in this language. On the other hand, the theorems are not trivial. To take a very simple example, one might think that if a and b cooperate, $B_a Pref_b(c, d)$ would imply $Pref_a(c, d)$. This is of course completely false, a and b can even when they have the same priorities have quite different beliefs about how the priorities apply to the constants. But the theorems show that no principles of this kind can be found that are valid only for cooperating agents. Moreover, they show that if one wants to prove that a formula like $B_a Pref_b(c, d) \rightarrow Pref_a(c, d)$ is not valid for cooperative agents a counterexample to it in which the agents do not cooperate suffices.

4.8 Conclusions and Future Work

In this paper, we have defined preference in terms of a priority sequence. In case agents only have incomplete information, beliefs are introduced. We have proposed three definitions to describe different procedures agents may follow to get a preference relation using the incomplete information. Changes of preference are explored w.r.t. their sources: changes of the priority sequence, and changes in beliefs. The multi-agent case has been investigated as well. For further study, we are aware that a large amount of research on preference has been done in social choice theory and computer science, we would like to compare our approach with this work. As mentioned earlier other types of priority are used in such research, often with weights. We do think our methods are applicable quite generally. Also, if only for comparison's sake, we will study preference between states (or propositions). Finally, preference is a key notion in game theory, we would like to see how our framework can be applied there.

Acknowledgement We thank Johan van Benthem, Reinhard Blutner, Ulle Endriss, Jerome Lang, Teresita Mijangos, Floris Roelofsen, Tomoyuki Yamada, Henk Zeevat, and two anonymous reviewers from the ESSLLI *Workshop on Rationality and Knowledge* in Malaga 2006 for their comments on previous versions of this paper. We thank the organizers of the *Workshop on Models of Preference Change* in Berlin 2006, and editors of this volume as well, Till Grüne-Yanoff and Sven Ove Hansson, for allowing us to present our work in the workshop. We thank two anonymous reviewers of this volume for their helpful comments. Finally, we thank Nina Gierasimczuk for her help with the pictures at the final stage of finishing the paper.

References

[ARS95] H. Andreka, M. Ryan, and P. Schobbens. Operators and laws for combining preference relations. In *Information Systems: Correctness and Reusability (Selected Papers)*. World Publishing Co, Singapore, 1995.

[Ben06] J. van Benthem. 'One is a lonely number': On the logic of communication. In P. Koepke Z. Chatzidakis and W. Pohlers, editors, *Logic Collquium, ASL Lecture Notes in Logic 27*. AMS Publications, Providence (R.I.), 2006. Research Report, PP-2002-27, ILLC, University of Amsterdam.

[Ben07] J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logic*, 17(2):129–156, 2007. Research Report, PP-2006-11, ILLC, University of Amsterdam.

[BL04] J. van Benthem and F. Liu. Diversity of logical agents in games. *Philosophia Scientiae*, 8(2):163–178, 2004.

[BL07] J. van Benthem and F. Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logic*, 17(2):157–182, 2007.

[BRG07] J. van Benthem, O. Roy, and P. Girard. Everything else being equal: a modal logic approach to ceteris paribus preferences. Research Reports, PP-2007-09, ILLC, University of Amsterdam, 2007.

[BS06] A. Baltag and S. Smets. Dynamic belief revision over multi-agent plausibility models. In *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory (LOFT 06)*, Liverpool, 2006.

[CMLLM04] S. Coste-Marquis, J. Lang, P. Liberatore, and P. Marquis. Expressive power and succinctness of propositional languages for preference representation. In *Proc. 9th International Conference on Principles of Knowledge Representation and Reasoning (KR-2004)*. AAAI Press Whistler, Canada, 2004.

[Cre71] M. J. Cresswell. A semantics for a logic of 'better'. *Logique et Analyse*, 14:775–782, 1971.

[DW94] J. Doyle and M. P. Wellman. Representing preferences as ceteris paribus comparatives. *Working Notes of the AAAL Symposium on Decision-Theoretic Planning*, 1994.

[FHMV95] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press Cambridge, MA, 1995.

[Gro91] B. N. Grosof. Generalising prioritization. In J. Allen and E. Sandewall, editor, *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pp 289–300. Morgan Kaufmann, 1991.

[Hal57] S. Halldén. *On the Logic of "better"*. Gleerup, Lund, 1957.

[Han95] S. O. Hansson. Changes in preference. *Theory and Decision*, 38:1–28, 1995.

[Han01] S. O. Hansson. *Preference Logic*, volume 4 of *Handbook of Philosophical Logic*, chapter 4, pp 319–393. Kluwer, Dordrecht, 2001.

[Jen67] R. E. Jennings. Preference and choice as logical correlates. *Mind*, 76:556–567, 1967.

[Kra81] A. Kratzer. Partition and revision: the semantics of counterfactuals. *Journal of Philosophical Logic*, 10:201–216, 1981.

[Lew73] D. Lewis. *Counterfactuals*. Blackwell, Oxford, 1973.

[Lew81] D. Lewis. Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic*, 10:217–234, 1981.

[Liu08] F. Liu. *Changing for the Better: Preference Dynamics and Agent Diversity*. Ph.D. thesis, ILLC, University of Amsterdam, 2008.

[Liu09] F. Liu. Diversity of agents and their interaction. *Journal of Logic, Language and information*, 18(1):23–53, 2009.

[MvdH95] J.-J.Ch. Meyer and W. van der Hoek. *Epistemic Logic for Computer Science and Artificial Intelligence*. Number 41 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge, 1995.

[PS93] A. Prince and P. Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell, Malden, M, 1993.

[Rot03] H. Rott. Basic entrenchment. *Studia Logica*, 73:257–280, 2003.

[Rot06] H. Rott. Shifting priorities: simple representations for 27 iterated theory change operators. In H. Langerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, pp 359–384. Uppsala Philosophical Studies 53, 2006.

[Tra85] R. W. Trapp. Utility theory and preference logic. *Erkenntnis*, 22:301–339, 1985.

[Vel96] F. Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261, 1996.

[Wri63] G. H. von Wright. *The Logic of Preference*. Edinburgh Universtity Press, 1963.

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):

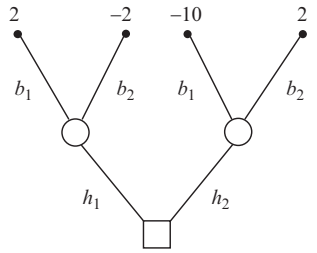

Fig. 5.1 Subtree T_1

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):

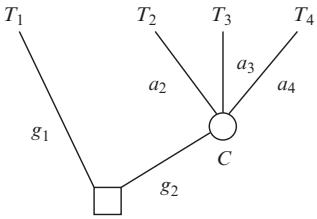

Fig. 5.2 Global model T_0

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

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

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

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

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 b_1 and b_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, b_1 and b_2, and prefers b_1 over b_2; say, $U_0(b_1) = 1$ and $U_0(b_2) = 0$, 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 b_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 $U_3(b_1) = 0$ and $U_3(b_2) = 1$. Thus, we have the following global decision model (Fig. 5.3):

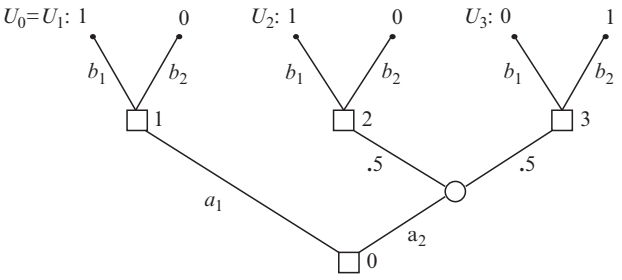

Fig. 5.3 The second global model

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.

Chapter 6
Exploitable Preference Changes

Edward F. McClennen

Abstract There is an extensive literature on the price that one can end up paying if one's choices do not satisfy certain axioms, e.g., Independence of Irrelevant Alternatives (IIA), and Independence (IND). The argument is that one can be turned into a "money pump," by another person, in which one will be repeatedly offered a sequence of choices for some small amount of money, but while one will prefer to accept the offer, it will yield no gain whatsoever. That is, on each round one will pay out a small amount of money and receive nothing in return. It is customary to suppose that this provides one with a solid and thoroughly pragmatic argument for retaining the axioms in question. Being turned into a money pump, however, presupposes that in the context of the offers that will be made, one reasons in accordance with backward induction. I argue that in the context of such offers the appeal to backward induction is simply unconvincing. That is, there is no reason to suppose that the conclusions of backward induction in such cases are at all relevant. This, in turn, implies that the dropping of either of the axioms in question does not really pose any pragmatic problem for a rational decision-maker. I close by reflecting on whether there are perhaps other cases in which backward induction is questionable.

6.1 Preference Changes in General

There are many different situations in which one can experience a change in one's preferences. One may have started to execute a plan that one judged to be best, given the information available, only to subsequently receive new information that leads one to no longer prefer that plan. Alternatively, one may find that the development of more sophisticated tastes results in a change of preference, or that a change is due to sheer boredom with one's usual way of proceeding. Then there is the case of addiction. Deciding in a quiet, thoughtful moment early in the morning not to smoke that day, one discovers that later that same day one just must have a cigarette.

E.F. McClennen
Professor of Political Philosophy, Syracuse University
e-mail: efmcclen@syr.edu

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

Addiction cases themselves vary. Ulysses has himself tied to the mast, and stops up the ears of his crewmembers, in order to prevent himself from turning the ship towards the Siren's island, believing their song to be irresistible. At a more prosaic level, one may decide not to take the shortest way home, there being an ice cream shop on that route, which will tempt one to stop and indulge oneself, something that one now does not want to do.

6.2 Exploitable Preference Changes

It is unlikely that there is some account that can be given that fits all these different cases. What I propose to do here instead is limit myself to a very distinct type of preference change, one that lies at the heart of an important argument that was first put forward by Davidson, McKinsey and Suppes, in defense of a certain principle or axiom of decision theory.[1] By way of illustration, suppose a person were to use a method for evaluating various alternatives that failed to satisfy the Independence of Irrelevant Alternatives (IIA) condition. According to this condition, among other things, if one strictly prefers x to y, when the options available are just x and y, then one should still strictly prefer x to y when some third alternative z is also available. This is not a case where one's preferences just happen to change. Rather it is a case where it is now and at all future times true that one, say, prefers y from the available set $\{x, y, z\}$ and x from $\{x, y\}$. Savage once proposed a rule for evaluating completely uncertain prospects – the minimax regret rule – that didn't satisfy this condition.[2]

The standard argument is that such a violation of IIA places one in a position of being turned into a "money pump." Davidson et al. (1955) originally used the money pump argument to rule out *cycles* in preference. But it is equally applicable to violations of the IIA condition, since those violations can generate cycles. To be sure, Savage's minimax regret principle generates well-behaved orderings over any particular set of alternatives – that is, orderings that are connected, fully transitive, and have no cycles – but if one alters the elements in the available set by adding or subtracting certain alternatives, a different and acyclic preference ordering may emerge. In short, the ordering of any x and y is not *context free*: the ordering can depend on whatever other alternatives are available.[3]

The problem of cycles that arises in the case of a violation of IIA can be illustrated by the following very simple situation. Suppose one violates IIA and prefers y when the available set is $\{x, y, z\}$ but prefers x when the available set is $\{x, y\}$. Suppose, further, that one is offered x at a price that one likes and so one accepts the offer. (What that price is plays no role in the argument.) Suppose finally that another agent secures y and z and then offers to let one either trade x and a small amount of

[1] Davidson et al. (1955).

[2] See, for example, Luce and Raiffa (1957), Chapter 13.

[3] See McClennen (1990), Chapter 2.

money, $e, for y, or trade x for z. Since one strictly prefers y to x, when all three are available, one should be willing to make the trade, so long as $e is small enough that $y - \$e$ is still preferred to x. At the end of this transaction, then, one has y, and has expended a small amount of money, $\$e$. The other agent could then offer one a choice between just x and y, and since in this case one prefers x over y, one should be willing to trade y and another small amount of money, $\$e$, for x. Thus one ends up where one started, except that one is out $\$2e$. Recast in extensive or tree form the problem looks like this (Fig. 6.1):

What makes the term "money pump" appropriate here is that the agent can be repeatedly offered these options until the violator of IIA runs out of money. It can also be argued that the encounter between the violator of IIA and some other manipulative agent is fully to be expected. As you repeatedly lose small amounts of money, the exploiter is gaining them, and given human nature to be what it is, one surely must expect that, even if friends and relatives would not pump you in this way (except perhaps your own older brother), there will be no lack of others who will be eager to enter into such exchanges. The moral of the story, of course, is that one's preferences must satisfy the IIA condition, if one is to avoid being pumped in this way. Notice that there is no need to appeal to special intuitions here in support of IIA: one has a perfectly "pragmatic" argument for not abandoning this condition.

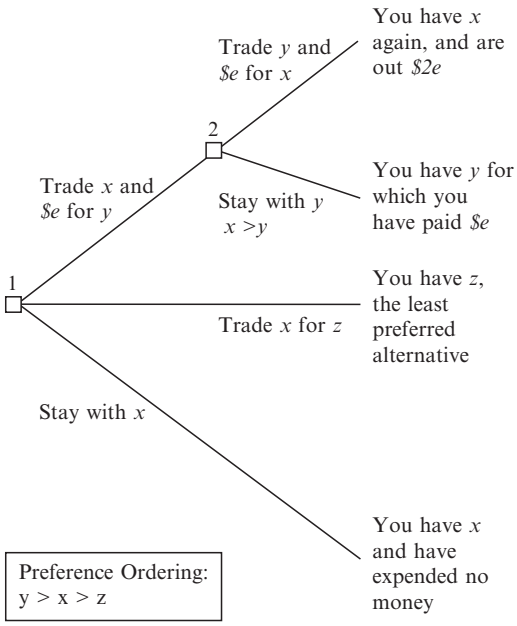

Fig. 6.1 The money pump argument for violations of IIA

6.3 Violations of Independence

A parallel argument can be constructed against the dropping of what is known as the Independence Axiom.[4] That axiom, in one of its more general formulations, requires that if g, g^*, and g' are any three gambles, and $g\ R\ g^*$, then $[g, p; g', 1-p]$ R $[g^*, p; g', 1-p]$, where p is a probability value, and R is the weak-ordering relation ('as good as'). In verbal form, if one gamble is at least as good as another, then compounding each with the same third gamble should not change the ordering. Consider now, for example, the following two pairs of alternatives:

$$g_1 = [\$2, 400, 1]$$
$$g_2 = [\$2, 500, 33/34; \$0, 1/34]$$
$$g_3 = [\$2, 400, 34/100; \$0, 66/100]$$
$$g_4 = (\$2, 500, 33/100; \$0, 67/100]$$

Suppose, for example, you would rank gamble g_1 preferred to g_2 but would rank g_4 preferred to g_3. You offer, by way of explanation, that the risklessness of g_1 makes it more attractive than g_2, but, when risk is unavoidable, the higher possible payoff of g_4 makes it more attractive than g_3. With a bit of algebraic fiddling, however, it can be shown that such a pair of preferences violates the Independence Axiom. Now consider the following problem in extensive tree form (Fig. 6.2):

If $\$e$ is small enough, we will have that $g_4 - \$e > g_3$. Now suppose that you begin by possessing g_3 (the exploiter has sold it to you at an acceptable price) and the exploiter now offers you an additional option, namely, g_4, at a price of $\$e$. You have to decide whether to retain g_3 or trade it in order to be exposed to the additional option of g_4. One will presumably accept the trade, for this gives one the opportunity to secure gamble g_4, which one prefers to g_3. Armed with that opportunity one now proceeds upwards at the first choice point, planning, if one reaches the second choice point, to choose upwards again, so as to expose oneself, in sequential form, to the gamble g_4. But if and when one gets to the second choice point, what one then faces directly is a choice between g_1 and g_2, and by hypothesis one prefers the former to the latter. Thus, so the argument goes, one will choose g_1. In effect, then, one ends up exposing oneself to g_3 rather than g_4, and loosing $\$e$ regardless of whether E or –E occurs, and this in comparison to what one would have received if one had simply stayed with g_3. The problem, then, stems from planning to choose g_2 at the second choice point, but then, if and when one does arrive at the second choice point (if event –E occurs, of course, one will not get to the second choice point) choosing g_1 instead.

As in the case of violations of IIA, the exploiter can repeatedly capitalize upon such a preference shift at the second choice point by getting the agent to buy g_3 (at

[4] This is also related to what is known as the Dutch-book argument. See, for example, Ramsey (1931). For the parallel between the independence of irrelevant alternative condition and the independence axiom for choice under conditions of risk, see McClennen (1990).

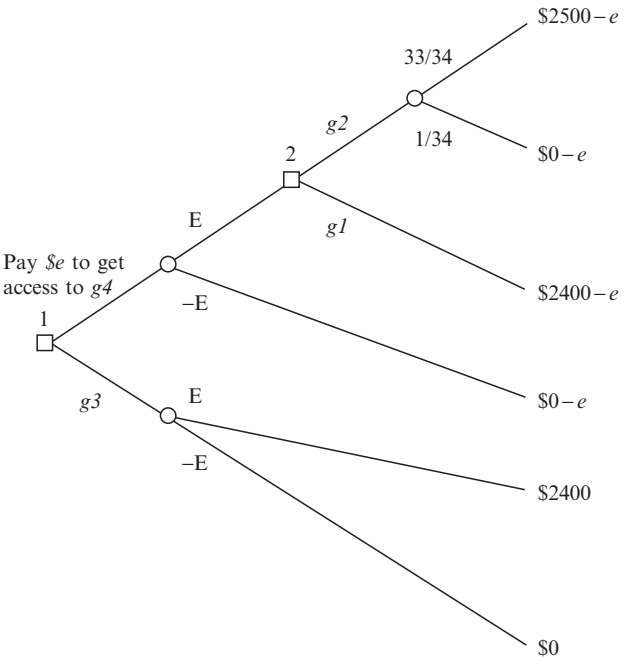

Fig. 6.2 The money pump argument for violations of Independence

whatever price he agrees to), and then getting the agent to pay $\$e$ for a chance to get g_4. At the first choice point it is g_4 the agent prefers, and so he or she is willing to pay $\$e$ to get g_4, but subsequently, upon arriving at the second choice point, the agent will presumably choose the option there that, coupled with the choice at the first choice point, exposes the agent to g_3, rather than g_4. The other player gets $\$e$, and the agent loses $\$e$, each time the agent is offered the option g_3, and then pays the exploiter $\$e$ and g_3 to get a chance at g_4. Moreover, once again, one can expect that exploiters will be there to take advantage of you.

These are the kind of preference changes that I want to examine in this paper: changes that can be exploited to the detriment of the decision-maker. In what is to follow, I shall refer to this kind of change as an "exploitable preference change." I think it would be odd to lump this sort of preference change with changes due to the receipt of new information, or changes that reflect temptations, minor (in the case of ice cream parlors), or major in the case of a Ulysses situation. At the very outset, one presumably knows that one will confront a choice between x, y, and z, and then subsequently one will have to choose again, this time, between just x and y, and similarly, confronting g_3 and g_4, and then (if one does arrive at the second choice point) subsequently confronting g_1 and g_2. What makes this kind of situation interesting is that one can be exploited, and ordinarily this will not be the case when one comes to have new information, or simple changes in taste take place. Similarly, a case of an exploitable preference change seems very different from one in which one is, in some sense, *driven* by cravings or temptations that are not easily controlled by the use of one's reason.

6.4 Myopia

How does one, as a violator of either of these axioms, get into such an unfortunate situation? How could it happen that one does not realize the trap to which one is exposed by having a preference ordering that does not satisfy IIA or Independence? The answer many have suggested is not that one has such preferences, but simply that one fails to look ahead – exercise foresight – and see what is coming. In effect, the problem arises because one is *myopic*. At each choice point one concentrates just on what is available at that point, and disregards the implications of where one is in a projected sequence of choices that are to be made. Becoming a money pump, then, is the fate of an agent who not only violates IIA, but who treats each choice point in a projected decision-tree as if it presented an isolated or *de novo* choice to be made.

6.5 The Sophisticated Approach to Exploitable Preference Changes

In the decision-theory literature one finds a standard suggestion, however, as to how a rational person can deal with such situations. This is to anticipate that the situation can occur, and take precaution in one way or another, if anyone confronts you with such a choice problem. In the case of violations of both IIA and Independence, what is to be anticipated is that one may be offered two trades in sequence, and recognizing that if one were to accept both, one would end up worse than one was at the start, one sees that a sensible thing to do is to refuse both exchanges, and simply stick with x. From this perspective it is only the failure to trace out the full implications of the situation – to myopically accept each trade as it comes along, not looking ahead to see what the final result of one's choices will be – that poses the problem. This way of resolving the problem is known as choosing in a *sophisticated* rather than a myopic manner.[5] One anticipates that if one accepts the first trade, and so trades x for y, and a small amount of money, $\$e$, and is then offered the second trade – to trade y and a small amount of money, $\$e$, to get x again – one will prefer to trade again, and thus will be worse off for the sequence of trades than if one refused at the outset to trade. Thus, one decides instead to stay with x. Similarly, in the case of violations of Independence, recognizing what will happen, one should refuse, having purchased g_3, to now trade it for a chance at g_4. Being rational, then, involves among other things, exercising *foresight*. Of course, at the first choice point, one would prefer to make one trade only, and so end up with y, for which one has paid $\$e$. But one judges that this option is not available. That is, one must anticipate that at the second choice point, one's preference will be to trade again. And similarly one must anticipate in the case of the gambles that one will choose g_1 and not g_2.

[5] See Strotz (1955) who calls this the strategy of "consistent planning."

In effect, one must reason *backwards* from the last choice point (or points if there is more than one) and expect at the first choice point that it will be rational to choose to trade again at the second choice point.[6] The way to deal with the problem, then, is to not trade at the earlier choice point. This would mean that there would be no second choice point, and in this way the cycles, x to y to x, and from g_3 to g_4 and then back to g_3 would be broken.

6.6 Implications for the Foundations of Decision Theory

Does the possibility of taking a sophisticated approach defuse the money pump argument?[7] It would seem that the sophisticated approach allows one to violate IIA and/or Independence and still avoid being pumped. In each case the problem can be avoided. It can be avoided simply by anticipating what one will do if one were to get to the second choice point, and acting in the light of this anticipation.

This in turn implies that the money pump argument does not provide a pragmatic defense of IIA or Independence. This might pose no problem if there were other convincing arguments in favor of the axioms in question. My own view is that there are no other arguments. I have recently set out my objections to the other arguments that have been offered in the case of Independence.[8] Defenses of IIA are complicated by the tendency of most to suppose that what is at issue is some sort of requirement that there be no cycles. I have no trouble with acyclicity, if what is being ruled out is some sort of cycle within the ordering of a given set of alternatives. But this is simply because such a cycle is simply incoherent. It violates the very notion of an ordering. What one should do is insist that acyclicity is simply true *by definition* of an ordering. If that is true, however one doesn't need to appeal to the money pump argument. Acyclicity is simply a non-starter. But there is no definition of an ordering that ensures that IIA is trivially true. Savage's minimax regret rule makes this clear. It always generates an ordering over any given set of alternatives that is fully transitive, connected, etc. It is just that if the set of alternatives is altered, the ordering may change. Similar results are to be found among certain voting rules such, for example, the Borda Count Rule. Without the money-pump argument, then, we are reduced to arguing unconvincingly that IIA is intuitively true. But none, as far as I know, has ever argued in that way for IIA.[9]

Perhaps one could argue that there is an alternative defense of conditions like IIA and Independence. Suppose the options in question all have the same relevant characteristics *across* choice sets, so that, for example, the chooser is indifferent

[6] This involves using the backward induction method of reasoning, about which I will have more to say below.

[7] To the best of my knowledge it is Schick (1986) who first argues this.

[8] See McClennen (2008).

[9] Weakening IIA generates a very interesting alternative to Nash's bargaining model as Kalai and Smorodinsky (1975), Kalai (1983) have shown.

between ending up with x when chosen from $\{x, y, z\}$ and ending up with x when chosen from $\{x, y\}$, etc. If this is true, then it might seem that the agent who violates IIA does seem to be asserting a flat contradiction: that x is better than y *and* that y is better than x.[10] But if the *contexts* of the options that are being considered are relevant (as in the case of, for example, Savage's minimax regret criterion) then the chooser will not necessarily be indifferent between x-as-chosen-from-$\{x, y, z\}$ and x-as-chosen-from-$\{x, y\}$, since it may well be the case that the maximum regret associated with the first is different from the maximum regret associated with the second.

6.7 Rabinowicz's Argument

More recently, Rabinowicz has argued that a sophisticated approach to sequential choices does not guarantee that one cannot be pumped.[11] Here is Rabinowicz's example (Fig. 6.3):

The bold lines stand for the moves prescribed by backward induction. Here, as in the problem in Fig. 6.1, suppose that there is a cycle in the ordering of x, y and

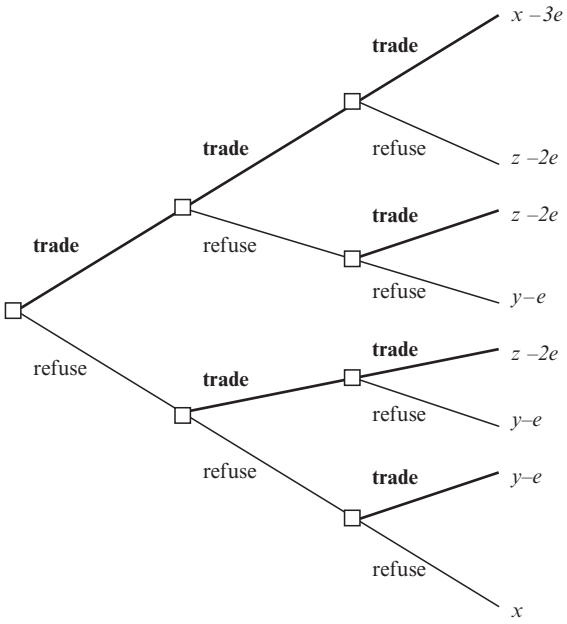

Fig. 6.3 Rabinowicz's argument

[10] I am indebted to an anonymous referee for suggesting this alternative.

[11] See Rabinowicz (2000).

z, so that $x < z < y < x$, and there is an amount of money e small enough such that $x < z - 2e < z < y - e < y < x - 3e < x$.

Appealing to backward induction, one should trade at each of the last choice nodes. Since one is aware of this conclusion, one should also trade at the two next-to-last nodes as well. But, then, one should also trade at the initial node. Trading at the initial node is predicted to lead to $x - 3e$, which is preferred to what one can predict will be the outcome, if one refuses to trade at the initial node, namely $z - 2e$. In this case, then, sophisticated choice does not ensure that one can avoid being pumped. Must we conclude, then, that standard decision theory can still rely on the money-pump argument and thus insist that axioms such as IIA and Independence must be respected?[12]

6.8 The Resolute Approach to Exploitable Preference Changes

As it turns out, there is another way in which money pump arguments against exploitable preferences can be defeated. Call this the *resolute* approach.[13] Consider again Fig. 6.1, where violation of IIA leads to a cycle. One could decide to accept the first trade, giving up x and a small amount of money, \$e, to get y, and then *refuse* the second trade in favor of staying with y. At the first choice point, where the options are x, y, and z, one prefers y to the other options, so one could decide to pay a small amount, \$$e$ to give up x and get y, but also plan to refuse the trade at the second choice point. One thus adopts a *plan* of trading x for y, at the earlier point, and then refusing, at the later point to exchange y for x. Despite the fact that at the second choice point one would, were one to view this from the perspective of backward induction, choose x over y, one could instead *resolutely* execute the plan calling for one to stay with y. Similarly, in the case of Fig. 6.2, where the cycle is generated by a violation of Independence, one could choose to spend \$$e$ for access to g_4, and then choose g_2 if and when one gets to the second choice point.

Both sophisticated and resolute choice have this in common: control is vested in the person at the earlier choice point. In the case of sophisticated choice, one takes action at the first choice point to ensure that the overall set of choices one makes does not leave one in a worse position than when one started. One does this by precluding certain subsequent choices from being made. The resolute chooser, on the other hand, settles upon a plan which calls for one, upon arriving at the second

[12] Rabinowicz (2000) offers a considerable more complicated conclusion, for he distinguishes between what he calls benign and vicious cycles, focuses on a wide range of cycles, and draws somewhat different conclusions than I have. My own analysis, e.g., in McClennen (1990), and in the present paper, is limited to a special set of cycles that arise as a result of violations of IIA and Independence. If I understand his discussion, the cycles I speak about are benign, that is, the choice sets upon which they are based are well-defined for the set of all alternatives, and all subsets thereof, and hence, to his way of thinking, they do not leave the agent liable to being pumped – see Rabinowicz (2000) – if the agent is resolute.

[13] McClennen (1990), 1.8

choice point, to stick with the y that one acquired at the first choice point, even though, looking at that second choice from the perspective of backward induction, it would seem that one would prefer to trade the y one now has for x and a small amount of money, \$$e$. Similarly, in Fig. 6.2, a resolute chooser would continue on to the second choice point and then resolutely choose g_2, in order thereby to execute the preferred plan.

How does resolute choice fare in the case of Rabinowicz's example? On the postulated ordering, x $< y - e < z -$ 2e, so the best plan is (trade, refuse, trade) *or* (refuse, trade, trade) – each having a predicted outcome of $z - 2e$, *not* $x -$ 3e. Thus, the resolute chooser escapes from being pumped.[14]

6.9 Against Resoluteness: The Argument from Backward Induction

The resolute chooser avoids being pumped, but it is clear that this approach, unlike the sophisticated approach, cannot be squared with the conclusions of backward induction. Indeed, it would seem that to defend resolute choice one must allow for the possibility of counter-preferential choice: in Fig. 6.1, the resolute chooser prefers at the later choice point to select x; but the resolute plan calls upon one to stand fast with y. That is, one must *refuse* at the second choice point to do what one would otherwise do, in the light of what are presumed to be one's preferences at the second choice point, namely to choose to exchange y and the small amount of money, \$$e$, for x, and to choose g_2 instead of g_1. In *Rationality and Dynamic Choice* I sought to rebut this by suggesting that the rational agent would, in virtue of preferring to execute the plan chosen at the first choice point, in fact prefer to continue that plan at the second choice point. The suggestion I made was that the preference for the plan in question would lead, in effect, to an "endogenous" preference change at the second choice point, and thus to an avoidance of the charge of advocating counter-preferential choice.[15] On this way of reasoning, the resolute chooser thinks in such a situation in a more holistic manner: he or she looks not just at each choice, as it presents itself, but looks at the whole sequence, selects the plan which – as defined by a sequence of choices – is most preferred, and then resolutely executes that plan. At this point, however, I confess that the introduction of what I called an "endogenous preference change" seems to me to be in need of much more discussion.

Let me begin by exploring the argument for why a rational agent must choose to trade back y for x at the second choice point in Fig. 6.1, and choose g_1 rather than g_2, at the second choice point in Fig. 6.2. If that argument is correct, then resolute choice at the second choice point is simply not feasible for a *rational* agent. The problem with defending resolute choice, of course, is that it runs afoul of the

[14] For the discussion of various variations on this ordering (Rabinowicz 2000).

[15] See McClennen (1990), 12.7.

principle of *backward induction*, to which appeal was made above in the discussion of the problems in Figs. 6.1 and 6.2. To rehearse this point again, backward induction requires that the choice made at any point in a decision tree must be consistent with the choice that would have been made if that point were the first choice node in a tree that began *de novo* there and that was identical to the subtree in the original decision tree. More specifically, the argument is that one can separate the subtrees consisting of the last choices the agent is to make, and determine how the agent would choose if that subtree were treated as a tree in its own right, and in this manner move backward in the tree, incorporating this information into the evaluation of what to do at earlier choice points. Thus, if in a decision tree consisting of just a choice between x and y, or just a choice between g_1 and g_2, *and* the rational agent would choose, respectively, x and g_1, then the rational agent must choose x, and g_1, at the second choice point in the full trees given in Figs. 6.1 and 6.2. But if that is true, then an endogenous preference change makes no sense (at least in this context) and resolute choice cannot be defended. The "if" clauses of both these claims must be conceded: I have acknowledged that in an isolated situation x is pairwise preferred to y, and g_1 is pairwise preferred to g_2. Then backward induction kicks in to require similar choices at the respective choice nodes in the original full trees in Figs. 6.1 and 6.2.

It is interesting to note that backward induction has typically been used in the case of sequential game theory, for example, where two players interact, each choosing in turn, but it clearly applies also to a single player making a sequence of choices. It was in that context, interestingly, that it was in fact first applied, and in that application it was known as Bellman's Principle of Optimality. Here is the standard definition:

> *Bellman's Principle*. The principle that an optimal sequence of decisions in a multistage decision process problem has the property that whatever the initial state and decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions.[16]

6.10 A Reply to This Objection

The argument from backward induction supposes that in the two sequential choice problems we have been considering we can isolate the part of the tree (the subtree) that begins at the second choice point, consider it as a separate tree, and reach a conclusion about what a rational agent would do in that (reduced) tree, and then plug all of this back into our analysis of the whole tree.[17] But is this transportation of the ordering in the *de novo* trees back to the full trees in Figs. 6.1 and 6.2 correct? At this point it would simply beg the question to say: yes, because that is what

[16] *McGraw-Hill Dictionary of Scientific and Technical Terms* (2003).

[17] Rabinowicz (2000), Section 6.4 rehearses various objections that have been made to backward induction, but the objection I shall raise here is, I believe, quite distinct from those that he discusses.

backward induction requires. At the very least we need to raise and try to answer the question whether backward induction applies to a case like this?

My view is that it does not. Granting that it makes sense to conclude that the agent will select x over y, and g_1 over g_2, when these are considered as isolated pairwise choices, it is not at all clear that this says *anything* about what it is rational to choose if and when one arrives at the second choice point in the whole trees as presented in Figs. 6.1 and 6.2. In the first of these examples, the choice is between (1) trading y back for x, when one has just traded x for y, at an additional cost of $\$e$, *and thus exposing oneself to being made into a money pump*, or (2) not trading y for x, and thus not exposing oneself to the risk of being pumped repeatedly. And similar considerations hold for the choices between the gambles. It is decidedly unclear that the subtree from the second choice point in either problem can be treated as equivalent to a *de novo* choice between trading y back for x and staying with y, and a *de novo* choice between g_1 and g_2. In each case, the separate choice and the embedded choice are simply two different choice problems altogether, and this precisely because the decisions to be made at the second choice point in each case is contextualized in a way that the *de novo* choice problems are not. The standard line of reasoning here, I suggest, is exactly what could be speciously taken as the correct analysis regarding the utilization of Savage's minimax regret principle: the fact that one would choose a over b when the option set consists of $\{a, b\}$ does not imply anything about what one would choose when the option set consists of $\{a, b, c\}$. The context makes a difference. So also in the cases I have been considering. The option x in the context of an option set consisting of $\{x, y\}$, is not the same as the option x in the context of the option set $\{x, y, z\}$. Moreover, the difference is clear from the nature of the problem. When at the second choice point, the selection of x has as its consequence not simply getting back x, but also setting oneself up for being made into a money pump, by paying out e but having x again. On the other hand, choosing x over y at the second choice point has no such consequence. Similar remarks hold for the choice between the gambles.

If one has foresight in the case of a sophisticated approach, then it is certainty also permissible to suppose that the resolute chooser has foresight. But given such foresight, the resolute agent could hardly be supposed to have "forgotten" that what is called for in the problem in question – *where the agent clearly wants to avoid being pumped* – is upon arriving at the second choice point, to refuse to choose to trade y back for x, etc. The concern that the agent at the first choice point has to avoid being pumped remains a concern at the second choice point. We cannot suppose that the agent merely *planned* at the first choice point to choose to stand pat at the second choice point with y and, correspondingly to choose g_2 over g_1. These plans still make perfectly good sense at the second choice point. Indeed, finishing up their execution *at the second choice point* is essential if one is to avoid being pumped. Thus, one cannot argue that what the agent decided to do at the first choice point is merely past history – a "sunk cost" that should not be taken into account in deciding what to do at the second choice point. Alternatively put, that the agent in fact decides to move on to the second choice point, with a view to stopping there and not trading y back for x, and choosing g_2 instead of g_1, is part of the *relevant*

context of the choice which he or she now faces at the second choice point. That context is highly relevant for the choices presented in the full trees, but is absent, and thus cannot play any role, in the *de novo* trees. What the agent would do, were he or she to face the *de novo* trees simply has no relevance for what should be done at the second choice point in the full trees. But to say this is plainly and simply to say that it is inappropriate to appeal to backward induction in the context of these problems of potential exploitation.

This line of argument, moreover, would appear to be correct in the case of *any* situation in which the agent is faced with being exploited for having preferences that do not satisfy IIA or Independence. But that means, in turn, that the charge that such preferences are exploitable collapses, and with it any argument to the effect that the possibility of being exploited can be used to defend the inclusion of IIA and Independence as axioms of rational choice. Recall also that I have been suggesting that the exploitation maneuver can be stopped by a resolute approach. If Rabinowicz is correct that one can construct cases in which a sophisticated response to attempts at exploitation can be thwarted – that money pumps can still be constructed – and if I am correct that resolute choosers cannot be exploited, then those who want to utilize decision procedures that violate IIA and Independence can do so with impunity, so long as they are prepared to be resolute.[18]

Does this argument – that the choice to be made at the second choice point must be contextualized – provide a general argument against the method of backward induction? I think not. I am not disputing the claim that in many choice situations one can plausibly use backward induction arguments. That is, I am not claiming here that the argument just presented works in any case in which the backward induction argument has been employed. What I want to argue here is simply that the backward induction is inapplicable in the case of money pump arguments, and this because in such cases one cannot suppose that what one does upon arriving at the second choice point must coincide with what one would do in the corresponding *de novo* trees. In the problems as presented, one is presumably still trying to avoid the possibility of being pumped when one arrives at the second choice point, but no such possibility occurs in the subtrees that can be defined as *de novo* decisions. That, in turn, leaves me unpersuaded that a method of evaluation that leads to violations of IIA or Independence can be rejected on the grounds that it exposes the user to the money pump argument.

6.11 Two Final Thoughts

It will occur to the thoughtful reader that in defending resolute choice by resisting the backward induction argument, one might also want to consider a similar defense

[18] Of course, Rabinowicz himself originally thought that being sophisticated prevented one from being exploited. See Rabinowicz (2000), 125. Let us hope that he does not, on further reflection, come up with an example that shows that even resolute choosers can be exploited!

of sophisticated choice. Couldn't sophisticated choosers insist that they no less than resolute choosers can show that backward induction is really not applicable to the reasoning of the agent who seeks to escape from being pumped. In my original discussion above, I assumed that backward induction did apply to sophisticated choice. Of course, in earlier discussions, it was thought that sophisticated choice could be used to get out of the money pump situation.[19] One point of Rabinowicz's article, "Money Pump with Foresight," however, was to show that the sophisticated chooser *could* be pumped. What if such a chooser were to insist, however, that backward induction does *not* constrain his or her choices, and this on precisely the same grounds that resolute choosers can use, namely that the relevant subsequent choices must be regarded as contextualized by the consideration that someone is trying to pump him or her? In this case, however, how does one distinguish between being resolute and being sophisticated?

I must also confess that reflection on the exploitable preference change case leads me to wonder whether backward induction in some other cases is so secure. Consider the centipede example that has been central to many analyses of backward induction.[20] If players were to contemplate the possibility of cooperating throughout the many steps in a typical centipede game, and if they were to get such cooperation underway at the beginning, it is no longer quite so obvious to me that they each must assume that there could be no cooperation on the 100th game (or whatever number the last game is), and that one is forced by that consideration and backward induction to infer that no cooperation is possible from the very beginning. Might not one want to insist that in the case of the 100th game, one cannot simply take it for granted that what a player would do if the 100th game were to be played *de novo* yields a straightforward answer to what a rational player would do in the case where that game is the last in a series of 100 games. There have been, of course, a number of other arguments that have been offered for not always accepting the backward induction argument. As best as I have been able to ascertain, however, none of them has any application to the cases I have considered. They have typically applied to interaction between two persons one or the other of whom might possibly have *beliefs* about whether the other player is rational. But if the choices are to be made by one person, what sense is to be made of a belief, say, on the part of that person at the first choice point about how rational he or she will be at the second choice point? Even more to the point, what sense can be made of the idea that at the second choice point the rational choice is to trade y for x, or to select g_1 instead of g_2? To do this is to simply assume that backward induction always holds, and that is to beg the very question that needs to be answered here. On the account that I have offered above, the agent's rational selection at the second choice point is to refuse to trade y back for x, and to refuse to select g_1, and this simply because this is the only way to avoid being pumped. Viewed in, perhaps the selection that the agent could make in other cases, such as the centipede problem, in order to sustain cooperation that is to the advantage of both players, is in fact the rational choice. Notice that the

[19] See McClennen (1990), 10.2 and Rabinowicz (1995).

[20] See, for example, Aumann (1998) and Brandenburger (2007).

force of thinking in terms of resolute choice is that the past, no less than the future, is relevant to the choice to be made in a sequence of choices. This means, among other things, that for the case of a known number of iterations, the last choice to be made in the sequence is not settled by noting that there is, in such a case, no shadow of the future. True enough, but the shadow of the past (the decision regarding what plan to execute) might still function to shape one's final choice.

References

Aumann, R. J. 1998. On the Centipede Game. *Games and Economic Behavior* 23: 97–105.

Brandenburger, A. 2007. The Power of Paradox: Some Recent Developments in Interactive Epistemology. *International Journal of Game Theory* 35: 465–492.

Davidson, D., McKinsey, J. C. C., and Patrick S. 1955. Outlines of a Formal Theory of Value, I. *Philosophy of Science* 22: 140–160.

Kalai, E. 1983. *Solutions to the Bargaining Problem*. Discussion Paper No. 556, Department of Managerial Economics and Decision Studies, Northwestern University.

Kalai, E. and Smorodinsky, M. 1975. Other Solutions to Nash's Bargaining Problem. *Econometrica* 43(3): 513–518.

Luce, R. D. and Raiffa, H. 1957. *Games and Decisions: Introduction and Critical Survey*. New York: Wiley.

McClennen, E. F. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. New York: Cambridge University Press.

McClennen, E. F. 2008. The Normative Status of the Independence Principle. In *The Oxford Handbook of Rational and Social Choice*, eds. P. Anand, P. Pattanaik, and C. Puppe. Oxford: Oxford University Press.

Rabinowicz, W. 1995. To Have One's Cake and Eat It, Too: Sequential Choice and Expected Utility Violations. *The Journal of Philosophy* 92: 586–620.

Rabinowicz, W. 2000. Money Pump with Foresight. In *Imperceptible Harms and Benefits*, ed. M. J. Almeida. Dordrecht, The Netherlands: Kluwer.

Ramsey, F. P. 1931. Truth and Probability. In *Foundations of Mathematics and Other Logical Essays*, ed. R. B. Braithwaite, 156–198. London: Routledge & Kegan Paul.

Schick, F. 1986. Dutch Bookies and Money Pumps. *The Journal of Philosophy* 83: 112–119.

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

Chapter 7
Recursive Self-prediction in Self-control and Its Failure

George Ainslie

Abstract The combination of human foresight and the discounting of delayed events in a hyperbolic curve is all that is needed to explain the learning of higher mental processes *from the bottom up*. These processes are selected by delayed rewards insofar as they counteract the over-valuation of imminent rewards that is also predicted by hyperbolic discounting. For instance, these processes come to interpret repeated, similar choices as moves in an intertemporal bargaining game resembling an iterated prisoner's dilemma. Perception of current choices as test cases for cooperation in such a game recruits the extra motivation experienced as willpower. Lines seen as criteria for such tests may be experienced as beliefs rather than resolutions. The chance that shifts of self-prediction may cause radical swings of motivation makes choice unpredictable from just knowing the person's prior incentives, even by the person herself; the resulting introspective uncertainty is arguably the subjective basis of freedom of will. A similar kind of recursive self-prediction explains how surges of emotion or appetite can be occasioned by symbols that convey no information about the availability of external rewards.

7.1 Introduction

There is a basic tendency for humans and nonhuman animals to change their preferences from larger, later (LL) rewards to smaller, sooner (SS) rewards in the absence of new information about their availability or proximity. This tendency is best called *impulsiveness*, although the term has also been used trivially to describe spontaneity or poor motor inhibition. I will first review work presented elsewhere on the hyperbolic shape of the function that describes devaluation of delayed reward: the problem that maintaining consistent choice poses for evolution, and how this shape is apt to govern both impulsive changes of preference and methods of limiting these changes. I will then expand on my previous suggestion that the most important of these methods, the interpretation of current choice as a predictor of

G. Ainslie
Research Psychiatrist, Veterans Affairs Medical Center, Coatesville
e-mail: George.Ainslie@va.gov

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

future choices, exemplifies a phenomenon that can be inferred not only in conscious impulse control, but in such basic experiences as freedom of will, emotion, appetite, belief, and character.

The observation of *recursive self-prediction* – self-prediction that is fed back to the ongoing choice process – is limited by its inaccessibility to controlled experiment, but this phenomenon is predictable from experiments that are not only controlled but precisely quantitative; and it can be tested by other, less direct means. In my view its existence challenges the conventional assumption that preferences govern only voluntary choices, and that preferences are in turn governed by an overarching faculty of will. It opens the possibility that a broader array of mental processes than is usually imagined competes in a common marketplace of reward, and that self-control and other higher mental functions can grow from the bottom up through interaction in this marketplace. Recursive self-prediction probably mediates a great deal of human experience.

7.2 Hyperbolic Discounting Poses a Problem in Adaptiveness

Impulsiveness is fully explained only by the finding that reward-seeking organisms devalue prospective events in a hyperbolic function (Ainslie 1975, 2001), which describes value as a simple inverse proportion of delay:

$$\text{Value} = \frac{\text{Value at no delay}}{[\text{Constant} + (\text{Impatience factor} \times \text{Delay})]} \tag{7.1}$$

Hyperbolic discounting raises the obvious question of how people ever avoid switching their preferences toward SS rewards as they come close – that is, achieve the consistent behavior that is the norm of rational choice theory (RCT; Herrnstein 1990; Boudon 1996) and the requisite for success in financial markets. This is not an issue for nonhuman animals, in which long range planning has been shaped by natural selection in the form of specific hardwired instincts. Animals mate, defend territory and hoard food for the winter not to ensure offspring, maximize resources, and prevent future starvation, but to gratify current urges. Even chimpanzees can wait only a few minutes to get increased amounts of favorite foods (Beran and Evans 2006). However, the necessity of coding long range rewards into lifelong instincts greatly limits a species' ability to learn new environmental contingencies. When an instinctive method of hoarding is cracked by interlopers, countermeasures will not appear for many generations if a new instinct has to evolve. It would clearly be more efficient for an organism to try different hoarding strategies on the basis of the long-term results they produce, so that failure would cause the loss of only the effort of a particular strategy, not a whole organism. There do exist examples where nature has given nonhumans an ability to learn from long-delayed consequences. In bait shyness, for instance, an animal can learn to avoid a taste that has been followed by sickness hours later, but the range of possible

learning is narrow: The cue has to be a taste rather than a visual appearance, and the consequence has to be nausea rather than somatic pain (Garcia et al. 1974). You might think that a mechanism of more flexible choice among outcomes of varied delays would have evolved much earlier; but the hyperbolic discount curves that move animals to promptly obey instincts make long range intertemporal choice potentially disastrous; SS rewards will tend to dominate LL ones. Given any ability to short-circuit the instinctual mating process, for instance, animals become vigorously autoerotic, as anyone who has visited the monkey house in a zoo can testify.

Hyperbolic discount curves have created a major pitfall for the evolution of flexible intelligence, to the extent that there is a serious question of how these curves evolved. There are two possible rationales, one of them unlikely. It could be argued that behaviors such as mating and fighting benefit the species at the expense of the individual's long range interest, so groups that discounted urges for them hyperbolically were selected; however, individuals' awareness of their long range interests evolved long after the form of the discount curve did. The more likely, and simpler, answer is that hyperbolic curves are a previously harmless manifestation of a universal psychophysical principle: that changes in a sensory quantity are perceived as a proportion of the baseline quantity – the Weber–Fechner law as applied to delay or some correlate of delay (Gibbon 1977). Such proportionality is also described by a hyperbola. Hyperbolic curves were harmless until organisms became intelligent enough to manipulate their sources of reward. As long as reward is controlled by the contingencies with which a species' instincts evolved, prompt obedience to those instincts will be the individual's best bet. Conversely, the hyperbolic shape may be what has limited the evolution of intelligence, but is so basic to the structure of motivation that it cannot be replaced at this late stage. Imaginative humans have learned to divorce pleasure from its original adaptive purposes to an enormous extent, mating, eating, and behaving in general to get pleasure rather than to increase reproductive fitness. Great skill at taming nature does not correlate (positively, at least) with the production of children in modern society. In combination with hyperbolic discounting, skill makes the individual dangerous even to herself. Control over reward lets her take her life in her hands, with enormous motivation to waste her resources – addiction is a human phenomenon. And when competing for these resources with an individual who has learned to evaluate them consistently over time – a human skill that I will discuss presently – she is at risk of becoming a money pump – someone who sells her winter coat every spring and buys it back at a higher price every fall (Cubitt and Sugden 2001).

The combination of intelligence and hyperbolic discounting clearly poses a risk, but one that some people seem to overcome fairly well. How does someone with hyperbolic discount curves sometimes manage to keep to the plans that her own foresight dictates? Furthermore, this question is not the greatest one posed by the hyperbolic discount function. Although motivational inconsistency is the first issue that comes to mind in contemplating hyperbolic curves, fundamental assumptions about the self come into question soon after.

7.3 Hyperbolic Discounting Creates Motivation for Developing Higher Mental Functions

The conventional idea of the self is that of a unitary executive that is entirely able to command some subordinate faculties – motor behavior, for instance, both current and future – and totally unable to control many other important processes such as appetites, emotions, and involuntary behaviors, especially the "negative" processes that would not be chosen deliberately. This self is substantial, impenetrable, and exempt from the strict laws of physical causality: It is felt to be substantial in the sense that it comprises more than the set of its motives, and has a form of inertia – the tendency of a choice to remain in place from the mere fact of having been made. It seems impenetrable in not being susceptible of analysis into simpler components. And although it can cause actions through its function of will, incompatibilist doctrines of free will state that it is not bound by causes acting upon it in turn (Clarke 2003). However, the hyperbolic shape of the basic discounting curve raises the question of whether any of this is necessarily so. Motivational theory can break free of the early behaviorists' model, the Skinner-box-writ-large that was so unlike the experience of complex choice (e.g. Skinner 1948), and contemplate higher mental functions with very different properties: held together only by motivation, analyzable with game theory, and predictive of the experience of free will while remaining strictly within the chain of causality as conventionally understood. If mental processes are shaped by a single, or common, selective factor that decays hyperbolically from the time of choice to the time of reward, it turns out to be fairly easy to model a self with these features.

Start with the concept of *value*, defined as the property of inducing behavioral selection: The functional effect of an event's value is the tendency of an organism to select a mental process that is followed by the valued event. A valued event is *a reward* (whereas the selective influence itself is just *reward*, without the definite article – potentially confusing, but it follows existing usage). The simplest model of choice is that an organism generates an array of options and selects the one that has the greatest expectable reward, discounted for delay and uncertainty. The precise way that options are generated and compared does not matter here, but it might be imagined to be something like Edward Tolman's concept of vicarious trial and error (1939), the rehearsal of each contemplated course of action before actually adopting one. Such a process has lately been observed physiologically in the rat hippocampus – the neurons subtending possible paths become active alternately until one path wins and choice moves forward (Johnson and Redish 2007). We would expect options that never win to eventually drop out of the array, so that reward affects not only the selection of a process but also the endurance of this process as an option.

If prospective reward were discounted in a function that produced consistent choice – *exponentially* – experience would affect subsequent choices only by changing the individual's expectations of delay and uncertainty. In a farsighted organism a faculty of self would be needed only to estimate what string of options chosen

consistently, would produce the greatest aggregate of discounted, expected reward over time. Selves would be mere calculators, and the process of choice would be determined by the estimated contingencies of reward, "throughput" as J. M. Russell called it (1978). Naturally theorists who imagine such a process of choice see the need to find extrinsic motives for impulsiveness such as sudden appetites driven by association, and for selves that perform impulse control by transcending mere motivation. However, given that prospective events are evaluated *hyperbolically*, options – or, more precisely, the mental processes that try to obtain these options – must compete with each other on the basis not only of each option's delay and uncertainty but also of their relative delays. Put another way, values shift relative to one another as a function of elapsing time, and thereby introduce an additional element of uncertainty to each option, even if the option is certain to be obtained if chosen. Mental processes that pursue contradictory options may each survive in an individual's array of choices because none dominates the others at all times. With a hyperbolic discount function, maximizing prospective discounted reward at one moment no longer "makes" a choice. To keep getting the reward that originally shaped it, the mental process pursuing that reward has to add means of staying chosen. The mind then functions as a population, not because it contains contradictory options – these would exist as well if rewards were discounted exponentially – but because the processes rewarded by these options have incentives to predict and forestall each other. This is the implication of hyperbolic discounting that lets it predict more than impulsiveness; it shapes the basic relationships that can ramify to form a self from the bottom up.

To reach fruition an option must promise not only the greatest discounted prospective reward of a current array of options *if it were certain*; it must also promise to withstand challenges by competing options that may look better before it comes to fruition. Its value is adjusted for the uncertainty that this very competition introduces. This problem can be demonstrated in, and sometimes solved by, a pigeon: If a peck on a red key leads to an SS reward, and no peck to an LL reward, and if an earlier peck on a green key simply keeps the red key from subsequently appearing, some birds learn to peck the green key (Ainslie 1974). This is impulse control of the simplest sort, and does not require the subject to have any functional knowledge of why pecking the green key leads to greater prospective discounted reward as of that moment. The pigeons that learn this kind of precommitment could be said to have foresight of a sort for the time periods involved – a matter of seconds – but not self-awareness. Even the most foresighted problem-solvers – people – have had limited success in devising impulse-control devices. External devices such as guardians and restricted bank accounts have limited availability and scope; diverting attention works only in the short run; and cultivating or inhibiting influential emotions (the psychoanalysts' reaction formation or isolation of affect) has significant costs. The external device that people have used most has been the influence of other people, sometimes in the form of physical controls – parents' control of their children, governments' enforcement of laws – but more robustly in other people's ability to give or withhold occasions for emotion (see Ainslie 1995). However, these social commitments also have limitations, especially as we devise

increasingly cosmopolitan societies. They become dangerous when you meet a person who wants to exploit you, a likelihood that increases with the number of people you meet; they give way when a whole group has the same impulse, a phenomenon that Jan Huizinga described as prevalent in the late middle ages (1924) but which still recurs in the form of "war fevers" and the "madness of crowds" generally; and they are useless against impulsive behaviors that can be concealed. The device that has best combined strength and flexibility has been another one altogether, which the individual exercises autonomously; it has been nebulous from the viewpoint of motivational science.

7.4 Recursive Self-prediction Provides a Mechanism for Will

An ability to stabilize one's own choice for one's own welfare was gradually differentiated from conscience in the sixteenth century, became a fashion in the seventeenth, and has been the subject of many theories since, often under the name of *will*. The early psychologists began cataloguing its properties (Sully 1884, pp. 630–670; James 1890, pp. 486–592), but the lack of externally observable markers led it to be stigmatized as an unscientific concept, and discussion of it dried up almost completely as the twentieth century unfolded (sketched in Ainslie 2001, p. 202, note 12). The will was held to be as inscrutable as the self (e.g. Pap 1961), from which it has not been clearly bounded. The absence of analytic discussion of a process that is so central to human functioning has been striking, suggesting a hesitation, even a queasiness, about putting mortal fingers on it, the kind of discomfort that some religions have had about naming their deity. However, from a scientific standpoint the main obstacle to analyzing the will has been the lack of a motivational rationale for it.

"Will" has been used to name the process by which intention is connected to motor movement, and the sense of ownership that someone has of her actions (Wegner 2002), but its most important meaning is the process that restrains impulses (See Ainslie 2004). The philosophers and psychologists who have given advice about the will over the centuries have discerned several attributes, most notably a basis in choosing according to principle rather than according to the particulars of the current circumstance. The power of this abstract idea to reduce actual impulsiveness is puzzling from the viewpoint of RCT, which depicts people as naturally consistent to begin with; but it is predicted by the hyperbolic discount function, given only two conditions: that the cumulative discounted value of a series of expected rewards is roughly additive, and that a person's expectation of getting the whole series can be made contingent on her current choice without physical commitment. The additivity condition has been verified experimentally (Mazur 2001; Kirby 2006), as has its implication that subjects will show greater preference for LL over SS rewards when choosing a whole series at once instead of singly. This increase in patience has been found in students choosing between amounts of money, and of pizza; subjects who chose every week for 5 weeks between a smaller, immediate amount and

a larger amount a week later were much more likely to choose the SS amount than subjects who had to make their choice for all 5 weeks at once, on the first week (Kirby and Guastello 2001). The same pattern has been observed in rats choosing amounts of sugar water (Ainslie and Monterosso 2003). The replication of this finding in animals shows that the increase in patience comes from the properties of the basic, presumably hardwired discount function itself, rather than depending on cultural suggestion or on an effect of total amount on patience (an effect seen only in humans – Green et al. 2004).

The second condition – that a person's mere perception of her current choice as a test case predicting how she will choose in the future can bundle series of choices together – does not lend itself to experimental test. However, the dependence of large expectations on current test cases is a common intuition. The cost to a dieter of eating a piece of chocolate is clearly not a detectable gain in weight, but her loss of the expectation that she will stick to her diet. Uncontrolled observations of several kinds support this intuition: The lore on willpower mentions a role for a bad precedent in reducing willpower (e.g. Bain 1859/1886, p. 440); when Kirby and Guasello suggested to their student subjects that each weekly choice predicted how they would make subsequent choices, they moderately increased the subjects' preference for LL alternatives (2001); and vulnerability to perceived lapses can be modeled by interpersonal bargaining games (Monterosso et al. 2002). However, the best way to test the original intuition is to sharpen it by a device popular in the philosophy of mind, the thought experiment. I have argued that a small number of selected thought experiments yield a valid rejection of the null hypothesis – that contingent self-prediction is unnecessary for volition (Ainslie 2001, pp. 125–139, and in press). Direct observation will be impractical for the foreseeable future; even functional magnetic imaging (fMRI), which has localized the components of many motivational processes (Cardinal 2006), cannot show the semantic content of such processes.

With our present observational abilities we can only follow out the implications of hyperbolic discounting, and test what we see against the familiar properties of volition: An individual with foresight who notices the predictiveness of present choices should develop processes that look very much like a will and a self by experience alone, without their being supplied *ex machina* by a homunculus: A self-aware hyperbolic discounter will learn to take into account the existence of other relevant processes that have been shaped differently by different temporal relations with the same reward center(s). Processes that are congenial to each other will cohere into the same process. Contradictory ones will treat each other as strategic enemies. Ineffective ones will cease to compete at all. Thus hyperbolically discounted reward will create what is in effect a population of reward-seeking processes that group themselves loosely into *interests* on the basis of common goals, just as economic interests arise in market economies. The choice-making self will have many of the properties of an economic marketplace, with a scarce resource – access to the individual's limited channel of behavior – bid for with a common currency – the prospect of reward. The logic of repetitive bargaining games will create regularities within this marketplace, including reliable support for those farsighted processes that can predict and act early to forestall or foster processes will be strongly motivated by

imminently available rewards. Maintenance and change of choice will be governed by *intertemporal bargaining*, the activity in which reward-seeking processes that share some goals (e.g. long term sobriety) but not others (when to have drinks) maximize their individual expected rewards, discounted hyperbolically to the current moment. This *limited warfare* relationship is familiar in interpersonal situations (Schelling 1960, pp. 21–80), where it often gives rise to "self-enforcing contracts" (Klein and Leffler 1981) such as nations' avoidance of using a nuclear weapon lest nuclear warfare become general. In inter*personal* bargaining, stability is achieved in the absence of an overarching government by the parties' recognition of repeated prisoner's dilemma incentives. In inter*temporal* bargaining *personal rules* arise through a similar recognition among the successive motivational states of an individual, with the difference that a future state is not motivated to retaliate, as it were, against past states that have defected. In the intertemporal case the risk of future states' loss of confidence in the success of the personal rule, and their consequent defection in their own short term interests, will present the same threat as the risk of actual retaliation. These contingencies can create a will without an organ, serving a self without a seat, just as the "will" of nations not to use nuclear weapons seems to be guided by an invisible hand.

In this way will can grow from the bottom up, through the selection of increasingly sophisticated processes by elementary motivations. In many depictions from Descartes onward the will has the appearance of a canoeist steering through rapids – using skill and foresight to ride forces much stronger than itself, but still something made of different stuff, a spirit, a homunculus. The intertemporal bargaining process grows the canoeist from the stuff of the rapids, different in skill and foresight but subject to the same motivational forces, and in fact developed by those forces. It is when the canoeist learns to include her own future tendencies as part of the currents she must anticipate that a pattern recognizable as a self develops. As with many natural patterns, this mechanism is most recognizable where pathology exaggerates it, for instance in obsessive–compulsive personality disorder and encapsulated areas of dyscontrol (Ainslie 2001, pp. 143–160). Here I will focus just on the way that recursive self-prediction permits the leap from current to canoeist, that is, from strict causality to the experience of free will.

When the incentives for alternative choices are closely balanced, small changes in the prospects for future cooperation swing the decision between cooperation and defection. In that case an assumption about the direction of the present choice will be a major factor in estimating future outcomes. But this estimate in turn affects the probability that the present choice will be in that direction. Thus the decision process is recursive – not tautological, but continuously fed back like the output of a transistor to its own input. If the person's predictions about her propensity to make the choice in question are at all open, this feedback process may play a bigger role in her decision than any given incentive, external or internal. For instance, a dieter faces a tempting food, guesses that she will be able to resist it, applies the consequences of this guess to the expected reward contingencies as an increase in the likelihood that she will reap the benefits of her diet, and thus has more to stake against the temptation. Then she discovers a credible loophole and thereby incurs

a fall in her expectation of a successful diet because of the chance she will try the loophole and not get away with it – that is, the chance that she will subsequently judge her choice to have been a lapse, thus reducing the stake against further lapses. This fall may be so great as to make the expected values of lapsing vs. trying to diet about equal, until some other consideration tips her self-prediction one way or the other. Such a process is not subtle conceptually, but it eludes any calculation based only on the contingencies of reward, and buffers the person's decision against coercion by these contingencies. Thus it can be argued to generate the experience of exercising free will (Ainslie 2001, pp. 129–134). Furthermore, such an explanation allows us to characterize free choices better than saying that they are too close to predict. After all, many behaviors are quite predictable in practice and are still experienced as free. What becomes crucial is the person's belief that a given choice depends on this self-prediction process, however she has come to represent this process to herself.

Diets and resolutions are examples of consciously constructed personal rules, with clearly defined conditions as to what kinds of choice are members of the relevant bundle, and criteria for which choices are cooperations and which are defections. However, once an individual has discovered that her current choice gives her predictive information about her future choices, even choices that are not governed by resolutions are apt to be influenced by this information to a greater or lesser extent. This influence will be largely nameless, or be hidden in seemingly disparate processes with names like force of habit, being true to yourself, or even responding to beliefs about the world. True, this recursive influence may sometimes serve purposes other than deterring impulses. For instance, I may habitually gather tasks to take to the office near my front door the day before I leave, either (1) so I can find them easily when I'm in a hurry, or (2) so as to keep myself from putting off doing them. Purpose (1) makes this activity a coordination game without a conflict of interest between myself currently and in the future; purpose (2) recognizes a repeated prisoner's dilemma, designed to coerce my future self by making any act of procrastination set a precedent. The difference may be perceptible in whether or not I experience the habit as having force: A coordination game can be changed without compunction if, say, a more convenient mnemonic device comes along. Change in a repeated prisoner's dilemma for what looks like momentary convenience may produce an unaccountable feeling of unease, which is a sign that I have suspected the choice was really a lapse of intertemporal cooperation.

7.5 Recursive Self-prediction Accounts for Sudden Appetites and Emotions

There is no reason why recursive self-prediction should be limited to conscious volition. There are many common experiences where a mental process that is under marginal control is influenced by signs of how it is progressing. J. M. Russell describes seasickness as an example:

> I suspect that I may be getting seasick so I follow someone's advice to "keep your eyes on the horizon".. . . The effort to look at the horizon will fail if it amounts to a token made in a spirit of desperation.. . . I must look at it in the way one would for reasons other than those of getting over nausea . . . not with the despair of "I must look at the horizon or else I shall be sick!" To become well I must pretend I am well. (1978, pp. 27–28)

Darwin said that emotions generally follow this pattern:

> The free expression by outward signs of an emotion intensifies it. On the other hand, the repression, as far as this is possible, of all outward signs softens our emotions. He who gives way to violent gestures will increase his rage; he who does not control the signs of fear will experience fear in greater degree. (1872/1979, p. 366)

Anxiously hovering over your own performance is common in behaviors that you recognize to be only marginally under voluntary control: summoning the courage to perform in public or face the enemy in battle, recall an elusive memory, sustain a penile erection, or, for men with enlarged prostates, void their bladders. William James went as far as to say that we feel an emotion only when we detect somatic manifestations of it – a theory that has been shown to be overstated (Rolls 2005, pp. 26–30), but which may well describe how quasi-voluntary processes are accelerated or modulated.

But how can processes that are more or less involuntary fit the same recursive pattern as will? The hyperbolic shape of the discount curve supplies an answer, by allowing us to broaden our concept of reward, and hence of motivation. The existence of an internal marketplace for positive incentives has long been assumed by utility theorists, economists foremost among them. Recently neurophysiologists have reiterated the necessity of recognizing such a marketplace (Shizgal and Conover 1996); that is, a mechanism by which all substitutable processes can be weighed against each other In a marketplace model many diverse processes compete for a limited channel of attention on the basis of a common dimension of selectability, such that an relative increase in this dimension for an act of game-playing, say, or charity, can lead it to be selected over an act of food consumption, while a relative decrease for the game or charity could lead the consumption to be selected. However, only desirable processes are usually imagined to compete directly with one another. Intuition has dictated that aversive processes participate only negatively in this marketplace – that they are introduced by a non-market process and have their effect only by making subsequent escapes rewarding. We use the words "reward" or "utility" for a property that is deliberately sought, and different words such as "urgency" or "vividness" for a property that seems to demand attention without being desirable, yet the latter terms also imply positive motivation – motivation that impels you into an experience. The notion that aversive processes are directly selectable along the same dimension as desirable ones seems to depart from intuition, but part of the problem is linguistic. If we stop equating rewardingness with desirability – the property that lets something be deliberately sought – and define it more basically as the property that makes whatever process it follows tend to be repeated, we can avoid having to explain the force of aversive experiences with a second, non-market process.

Examples such as nausea, rage, and fear are processes that are usually thought of as unmotivated – what is the incentive to be nauseated? – but rather imposed on the individual by a reward-independent process such as classical conditioning. An opposing view has long pointed out that the selective factors in classical conditioning – unconditioned stimuli – invariably have incentive value as well as the power to condition, and has suggested that conditioning is a form of reward-governed learning (Hilgard and Marquis 1940; Donahoe et al. 1993). The difficulty with this theory is that the incentive value of unconditioned stimuli is often negative, that is, that they select for processes which the individual is motivated to avoid. The frequent vividness of the negative emotions has seemed to demand a second kind of selective factor, which rewards attention while deterring physical approach. In the conventional model, pain, fear, grief, anger, and presumably nausea are imposed in reflex fashion either by innately programmed turnkeys or by stimuli that have been associated with such turnkeys. However, conditioned attention and reward-seeking participation look very much alike. The reward-responsiveness of negative emotions can sometimes be discerned in the cases where they have come under voluntary control: Sometimes people have learned to pay attention to a painful stimulus without emitting the emotion-like response that makes pain aversive ("protopathic" as opposed to "epicritic" pain – Sternbach 1968), or to withhold a fear response to stimuli that have been provoking it (Clum 1989). Anger may feel imposed by a circumstance, but everyone has sometimes experienced the competition between "bothering" with an anger and carrying on the activity that it threatens to spoil – a competition that is apt to turn on the rewardingness of the alternative activity. Indeed, anger shares many psychometric and neurophysiological properties with the more obviously positive emotions, such as increased optimism, heuristic as opposed to reflective cognitive processing, and left as opposed to right frontal cortical activation (Lerner and Tiedens 2006).

I have argued elsewhere that the hyperbolic discounting of reward permits the modeling of negative, positive, and mixed emotion-like processes by the cyclic mixture of reward and subsequent inhibition of reward (Ainslie 1992, pp. 100–114; 2005). To summarize briefly: Just as a cycle of binge and hangover attracts and then repels behavior over a period of days, and as nail-biting or tics attract choice only when they are possible within seconds (cf. Berridge's "wanted but not liked" behaviors, 2003; also Peciña et al. 2006), so an urge to panic or attend to a traumatic memory may be "satisfied" only for a split second before its aversive effect is felt. Such an urge attracts attention but deters physical approach, exactly the effect of conditioned *negative* emotions. For motivated *positive* emotions, the question is why they would not lead to autistic self-reward. The brief answer is that hyperbolically-based preference for SS over LL emotional experiences should motivate premature satiation unless this activity is limited to adequately rare occasions; I shall say more about this presently. Even daydreams must include obstacles if they are to escape complete habituation. Finally, the mixture in *mixed* emotions is not a weighing of two opposite valences – which would lead to neutrality – but rather the perception that a strongly motivated emotion will bring just enough aversiveness to make its desirability from a distance ambiguous.

The ability of negative incentives to compete in the internal marketplace on the basis of a single selective factor – reward – permits a wide range of involuntary processes to be brought into this marketplace. The set of reward-seeking behaviors will comprise all internal processes to the extent that they compete with one another for expression. In particular, emotion becomes a form of behavior. The sensation of being cut or burned offers an opportunity for the emotion of protopathic pain – an opportunity that is hard, but not necessarily impossible, to refuse. The sensation of tossing in a boat offers the opportunity for nausea, the perception of loss offers the opportunity for grief or anger, and so on. Many processes remain outside of this set, for instance the competition of a muscular extension reflex with an opposing contraction reflex; and many processes take part in the set only partially. Cardiac contractions and peristalsis are somewhat autonomous, in that they will occur regardless of that an individual is thinking or feeling, but regrets, daydreams, plans for dinner, awareness of an itch, and excruciating pain all compete with each other, however unequally. The more one occurs, the less room the others have to occur. Even cardiac contractions and peristalsis can be brought into this marketplace to a limited extent, when sensations from them come to attention or when activity in a market member (e.g. fear) raises or lowers their activity; but their core functioning remains outside the market. Sometimes a pathologic phenomenon shows that a seemingly autonomous activity must have been occupying a small space in the market, as when loss of the urge to breathe – a motivation not usually noticed – impairs respiration ("Ondine's curse"; Kuhn et al. 1999). Sometimes deliberate learning enlarges the market-responsive component of autonomous activities, as when cardiac contractions or peristalsis come under the control of hatha yoga or biofeedback (Basmajian et al. 1989). The boundaries of the internal marketplace are not sharp and may be variable to some extent, but they clearly include much more than the set of voluntary activities or the set of desirable activities. The point for the present discussion is that not only deliberate but also involuntary reward-seeking processes should be affected by recursive self-prediction.

The value of the marketplace model can be seen in the example of sudden craving. Conditioned appetite has been proposed as the explanation of the sudden cravings that people develop for food or drugs when they encounter reminders of them, particularly when the people are trying to avoid consuming them (Loewenstein 1999; Laibson 2001). However, in laboratory examples of conditioning, conditioned stimuli lead to responses only when they predict imminent consumption. If a conditioned stimulus (CS) occurs or begins well before its unconditioned stimulus (UCS) is due, subjects learn to estimate the delay and emit the conditioned response (CR) just before the UCS (Kehoe et al. 1989; Savastano et al. 1998; see Ainslie, 2009). The alternative that hyperbolic discounting makes possible is that appetites are reward-dependent processes, and that their sudden arousal in the absence of any increased availability of their objects is an attempt to make consumption of these objects more likely. The logic is as follows: Reward-dependent processes compete for acceptance on the basis of the current discounted value of the prospective reward for these processes. An appetite arises when an individual perceives the opportunity for consumption that can be made either more

rewarding or more likely by this appetite; appetite may serve not only to prepare for consumption, but to make consumption more likely. In examples of elicited appetite in the laboratory, the timing of consumption is necessarily controlled by the experimenter. In daily life, by contrast, goods that might be consumed impulsively are available much of the time, and their consumption is limited by a person's decisions. If a random appetite increases the rewardingness of a prospective object, it increases the likelihood that the person will consume the object, which will induce further appetite in preparation for the possible consumption. This is a positive feedback system, driven by the person's recursive self-perception of the likelihood that appetite will be enough to make her decide to consume the object. It has the same math as Russell's seasickness, the expectation of vomiting that confirms itself.

A sudden spike of appetite could thus come from the existence of positive feedback conditions. These conditions may obtain whenever the person's consumption is determined mainly by her choice about a readily available consumption good, but are apt to have the strongest effect when there is weak-to-moderate resolve not to consume: Where a person is not trying to restrain consumption she will keep appetite relatively satisfied; where she is confident of not consuming regardless of appetite she will not expect appetite to lead to consumption. In neither of these cases will appetite be rewarded by motivating consumption. In a recovering addict or restrained eater, by contrast, cues predicting that she might lapse could significantly increase the likelihood of lapsing. There will still be constraints on the motivation for an appetite – in modalities where unsatisfied appetite brings hunger pangs or withdrawal symptoms these will be deterrents; and appetite without a limited occasion will extinguish (see Ainslie 2001, pp. 166–171) – but the explosive appetite that so often ends people's efforts at controlled consumption can be understood as a motivated process that has sought to do exactly that.

This model depends on the hyperbolic shape of the discount curve, since an individual with consistent preferences over time would have no short range motive to undermine her own resolutions, or indeed any long range motive to make resolutions in the first place. Given such motives and some self-awareness, recursive self-prediction can be expected to punctuate consistent behavior with fits and starts of appetite.

7.6 Beliefs May Arise Through Recursive Self-prediction

In a model of the individual as a population of reward-dependent processes, facts can be seen as what constrains the search for reward. The experience of being constrained by facts is called belief. In highly imaginative organisms such as humans relatively little reward comes from current sensory experience, or even from the prospect of any sensory experience that is so imminent that it demands attention. Most of our significant prospects are relatively distant, complex, and subject to interpretation. These prospects reward us as occasions for current emotion, in competition with other occasions such as the vicarious experience of another person, or

pure fantasy ("make-believe"), as well as sensory experience itself. As I mentioned above, an occasion paces reward most effectively when it is relatively infrequent, which in practice means that it must be governed by contingencies other than the immediate rewarding potential of the emotion, and connected to the emotion in some way that lets it stand out from other possible occasions.[1] To keep from paling into a daydream the joy of winning must be occasioned by new information, specific to a person or project or sports team or even fictional story to whom or to which you have already given importance. Similarly an occasion for panic must have some connection to pain or loss, but will be less apt than joy to pale, because of your avoidance of such occasions.

Although there are many possible rationales for making occasions unique – a longstanding practice or a myth shared by an entire culture or even good fiction-writing technique – the simplest way of being unique is to be factual. The scenarios that are instrumental in changing the real world are apt to also be those that compete best in the marketplace at the current moment, but not necessarily because of the prospect of experiencing their practical results; they have hedonic impact beyond this prospect as occasions for emotion that are more unique than make-believe (see Lea and Webley 2006). However, the motivational impact of make-believe can be amplified to a comparable level by reducing the freedom to choose alternatives; commitment to the outcomes of particular fictional scenarios in online fantasy projects such as Second Life may yield emotions as imperative as "realistic" activities such as day trading. What makes Second Life more powerful than a video game is the extent to which it is a single consensual project that cannot be cheaply abandoned for another one. Fictional works may achieve this uniqueness by becoming cultural icons – as Schelling (1986) describes for the death of Lassie (1986) – or even by an individual's single-minded devotion to one immutable set of outcomes.[2] Such examples elevate "make believe" to made beliefs – commitments to occasions for emotion that are divorced from instrumental effectiveness in the real world but which are binding enough to have the same hedonic impact. If belief is basically the experience of being constrained by facts, the irreplaceable ingredient is not the descriptive truth of the facts but rather the emotional cost of escaping them.

The role of the perceived facts themselves is often unclear. We have a strong tendency to discern facts underlying constraints, but to the extent that practical instrumentality is not important, the facts that we identify may serve more as labels for particular constraints than as predictors of external rewards. Perhaps the most important source of these constraints that do not come from physical limitations is intertemporal bargaining. One example is the way that people experience the non-predictive cues that lead to appetites, described above. As with all processes for

[1] For aversive emotions these requirements are less stringent, since a person's motivated avoidance of them keeps them uncommon; also, for evolutionary reasons aversive emotions seem to habituate less than pleasurable emotions.

[2] A fictional but credible example is the hero of Robert Coover's *The Universal Baseball Association* (1968) who has invested his emotions so much in a single, long-continuing fantasy baseball game that the randomly determined outcomes have the impact of facts (Ainslie 1992, pp. 313–315).

which reward is freely available a cue is needed only to give occasion, that is, to select one moment from among many to make a focused bid for expression. Often the environment is a strong selective factor – coming upon food or a loss or a confrontation – but often the occasion comes from a mere reminder or symbol. Even then, a cue that leads to a feeling one time becomes more likely to do it the next time, because it increasingly stands out from other available occasions as the association is repeated. Soon it will be experienced as "the reason for" the appetite or emotion. That is, even when the first occasion was a random stimulus its evocativeness will come to seem like a fact of the external world.

Personal rules supply another important example of perceived factuality that comes from intertemporal bargaining. The very volatility of recursive self-prediction means that people will be apt to cling to rationales for truces, that is, to lines between do-able self-control and futile efforts. Again uniqueness is valuable – here the quality of being a *bright line*, a boundary between conflicting interests that cannot be shifted without inviting more shifts. A recovering alcoholic has an available bright line between some drinking and no drinking at all. A dieter has only lines laid down by diets, which are much dimmer in the sense that they are more replaceable by other authors' lines that do not stand out any less. Lines like these, which are the criteria of personal rules, are often experienced as facts, the more so the brighter they are. For instance, recovering alcoholics have long believed that they have a biological susceptibility that causes a single drink to lead to irresistible craving; but it has been shown experimentally that it is the belief that they have had a drink of alcohol, not the alcohol itself, that is followed by craving (Maisto et al. 1977).

Our inherited instinct for disgust turns upon mostly ambiguous stimuli in the modern world. The process of recursive self-prediction creates the belief that some things *are* dirty, occasions for disgust, and others *are* clean. Accepted authorities may alter boundaries between them, as when the mania for cleanliness early in the twentieth century followed the discovery of germs, and is said to be resurgent now after the discovery of new diseases (Ashenburg 2007, pp. 239–289); but often our instinct for disgust seems to be controlled by rituals of just sufficient difficulty, adjusted for how strong our individual instinct is to begin with. For instance, there is no scientific reason to avoid touching urine, your own or someone else's. Only a single, tropical disease is transmitted through human urine, and that not through simple contact.[3] Nevertheless urine is universally assumed to *be* a contaminant, a belief that waxes and wanes, however, inversely with the difficulty of avoiding it. Parents of young children experience a sudden reduction in their belief, and people on camping trips are not generally bothered by the impossibility of washing after urination. The British colonial army in the nineteenth century could carry only limited equipment on bivouac, and used the same trough for washing in the morning that they had used as a urinal the previous evening (Farwell 1985). Reduction in the behavior of urine-avoidance drives a reduction of the belief that it is needed, a change that is even more apparent in the converse situation of germ-phobics: avoidance of

[3] One strain of schistosomiasis, a parasitic infection, is spread by infected urine in bathing sites (Cox 1993).

a new kind of contact, with a doorknob, say, sets a precedent of treating doorknobs as contaminated, and is in danger of making the person grasp them only through a handkerchief in the future. The most effective treatment of this and other phobias is behavioral – graded exercises in which the patient acts as if the fear were not true (Marks 1997).[4] Of course the same person may to lip service to very different beliefs, but the actual constraint she is under is the behavioral boundary established by recursive self-prediction.

The belief that you have found a bargain can be instantly rewarding. The hunt for bargains produces the pleasure in many kinds of shopping, whether "compulsive" or not. However, maintenance of this belief requires behavior that is consistent with it. If you have stocked up on food at a good price or bought a concert series at a discount, you may face an incentive to eat the food when you are tired of it or attend a concert you do not expect to enjoy in order to avoid recognizing a loss. And yet you may be fully conscious of the unpleasant prospect. The belief in the bargain is really a personal rule for playing a game, the wins in which occasion emotional reward that is related only tangentially to the reward of tasting the food or listening to the concert. The relationship is that the prospect of this consumption authenticates the bargain-hunting as an instrumental activity rather than a mere game, even though, once so authenticated, the bargain-hunting is a self-sufficient source of reward and has requirements that sometimes contradict those of optimally consuming the ostensible reward.

Another personal rule that masquerades as a belief is a performer's self-confidence. A performer can be defined broadly as anyone whose activity can be ruined by a loss of nerve – comedian, acrobat, public speaker, even warrior or lover. The belief has the form, "I *am* able (funny, nimble, persuasive. . .)," but it depends on the behavior of not fleeing, literally or emotionally, from the activity. Such flight, incisively named "flopsweat" by comedians, has the same incentives as any other kind of panic – the insubstantial relief of gratifying an urge that nevertheless beckons insistently. A large component of the self-confidence is the expectation that you can avoid panic, which adds a stake to the avoidance but perversely, in this case, increases the urge to panic for that very reason; thus self-confidence is particularly prone to the positive feedback phenomenon. Performers often find that they need additional resolutions: avoiding defensiveness, not playing for applause, not copying past work, and other formulae for resisting short range rewards; these again may take the form of beliefs: "The audience doesn't matter" or "I'm doing this for art's sake."

The difference between a conscious resolution and a constraint that is experienced as a fact may sometimes lie in how much of your prospective reward is at stake in the relevant choices. Conversely, you may increase the prospect at stake and thus your motivation for self-control by interpreting your personal rule as a response dictated by a belief. A person who resolves to be vegetarian to conserve the earth's resources does not face a strong incentive never to backslide; a person who

[4] Compare Arnold Bennett's advice for curing "fussiness" by deliberately acting contrary to fussy beliefs about yourself as soon as you identify them (1918, p. 80).

believes that animals *are* fellow souls and eating them *is* murder will be committed much more strongly, to the point even that she will begin to experience disgust rather than pleasure at the thought of eating meat (see related studies by Paul Rozin, e.g. Rozin et al. 1997). A single lapse will have much broader implications than it would for the environmentalist, perhaps instilling doubt about her basic character.

An increased stake in a personal rule will increase the ease of following it, but also increase the loss if you do not. The increase in stake could come either from a long history of success, or the perception of this rule as a key component of a broader and more important rule – against cruelty, dishonesty, or perversion, for instance. At some point you will cease to perceive the rule as a resolution and experience it instead as a trait of your character: "I *am not the kind of person* who . . ." can kill, is sneaky or mean spirited, or might have a disgusting paraphilia. This is a stake that is threatened by even a single lapse, greatly increasing your motive to avoid catching yourself lapsing. It is arguably the maneuver discovered by John Calvin, which gave the early Protestant burghers their legendary ability to defer consumption (Weber 1904/1958; see also my discussion in Ainslie 2001, pp. 134–139): If any sin is a sign that you are among those predestined to damnation, it makes a sin much more important than just a single failure of good works. If you have such a belief, a lapse faces you with a choice among (1) modifying your belief, but thereby giving up its committing power; (2) accepting the prospect of damnation, which in motivational terms is probably the same as #1; or (3) rationalizing so as not to classify the behavior as a lapse – usually the least costly solution, and probably the greatest source hypocrisy where deceiving others is not a factor. Although it is conventional to distinguish character traits from behaviors that are merely habitual, and is thus natural to distinguish the "self-signaling" that will not tolerate lapses from less consequential self-prediction (Prelec and Bodner 2003), they are just different zones on a continuum.

7.7 Conclusions

Recursive self-prediction is the expectable consequence of hyperbolic discounting in self-aware individuals. It is inaccessible to controlled experimentation, but offers a parsimonious model of several otherwise puzzling human phenomena:

- Higher mental functions, exemplified by will, do not require unified faculties but rather can be seen as intertemporal bargaining skills that become included in reward-seeking mental processes to the extent that they lead these processes to be better rewarded.
- "Free will" describes the experience of predicting your choices in a way that also modifies these choices, making them unpredictable from a knowledge of the original incentives but not excepting them from literal causality.
- Involuntary processes such as appetite and emotion may be selected by the same mechanism that selects deliberate choices, the recursive prediction of which

explains sudden eruptions following "conditioned" stimuli, without our having to attribute special properties to the association process.

- Belief can be seen as the recognition of constraints on choice, which include incentives that are recruited through self-prediction but that are experienced as facts. The perception of commitment to some kinds of self-control as a character trait increases the extent of this commitment.

The inadequacy of previous bottom-up theories in explaining higher mental processes may have been due to their depiction of motivation as a linear product of the person's incentives. Recursive self-prediction is not an exceptional process, but is probably present in most human intentionality. To paraphrase physicist Stanislaw Ulam, "the study of non-linear motivation is like the study of non-elephant zoology."

References

Ainslie, G. 1974. Impulse control in pigeons. *Journal of the Experimental Analysis of Behavior* 21: 485–489.

Ainslie, G. 1975. Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin* 82: 463–496.

Ainslie, G. 1992. *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*. Cambridge: Cambridge University Press.

Ainslie, G. 1995. A utility-maximizing mechanism for vicarious reward: Comments on Julian Simon's "Interpersonal allocation continuous with intertemporal allocation". *Rationality and Society* 7: 393–403.

Ainslie, G. 2001. *Breakdown of Will*. Cambridge: Cambridge University Press.

Ainslie, G. 2004. The self is virtual, the will is not illusory. *Behavioral and Brain Sciences* 27: 659–660.

Ainslie, G. 2005. Précis of 'Breakdown of Will'. *Behavioral and Brain Sciences* 285: 635–673.

Ainslie, G. 2009. Hyperbolic discounting versus conditioning and framing as the core process in addictions and other impulses. In *What Is Addiction?*, eds. D. Ross, H. Kincaid, D. Spurrett, and P. Collins. Cambridge, MA: MIT Press.

Ainslie, G. and Monterosso, J. 2003. Building blocks of self-control: Increased tolerance for delay with bundled rewards. *Journal of the Experimental Analysis of Behavior* 79: 83–94.

Ashenburg, K. 2007. *The Dirt on Clean: An Unsanitized history*. Knopf.

Bain, A. 1859/1886. *The Emotions and the Will*. New York: Appleton.

Basmajian, J. V. 1989. *Biofeedback: Principles and Practice for Clinicians*. 3rd Edition. Baltimore, MD: Williams & Wilkins.

Bennett, A. 1918. *Self and Self-Management*. New York: George H. Doran.

Beran, M. J. and Evans, T. A. 2006. Maintenance of delay of gratification by four chimpanzees (Pan troglodytes): The effects of delayed reward visibility, experimenter presence, and extended delay intervals. *Behavioural Processes* 73: 315–324.

Berridge, K. C. 2003. Pleasures of the brain. *Brain and Cognition* 52: 106–128.

Boudon, R. 1996. The "rational choice model:" A particular case of the "cognitive model." *Rationality and Society* 8: 123–150.

Cardinal, R. N. 2006. Neural systems implicated in delayed and probabilistic reinforcement. *Neural Networks* 19(8): 1277–1301.

Clarke, R. 2003. *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.

Clum, G. A. 1989. Psychological interventions vs. drugs in the treatment of panic. *Behavior Therapy* 20: 429–457.

Cox, F. E. G. 1993. *Modern Parasitology: A Textbook of Parasitology*. 2nd Edition. Hoboken, NJ: Wiley.
Cubitt, R. P. and Sugden, R. 2001. On money pumps. *Games and Economic Behavior* 37: 121–160.
Darwin, C. 1872/1979. *The Expressions of Emotions in Man and Animals*. London: Julan Friedman.
Donahoe, J. W., Burgos, J. E., and Palmer, D. C. 1993. A selectionist approach to reinforcement. *Journal of the Experimental Analysis of Behavior* 60: 17–40.
Farwell, B. 1985. *Queen Victoria's Little Wars*. New York: Norton.
Garcia, J., Hankins, W., and Rusiniak, K. 1974. Behavioral regulation in the milieu interne in man and rat. *Science* 185: 824–831.
Gibbon, J. 1977. Scalar expectancy theory and Webers law in animal timing. *Psychological Review* 84: 279–325.
Green, L., Myerson, J., Holt, D. D., Slevin, J. R., and Estle, S. J. 2004. Discounting of delayed food rewards in pigeons and rats: Is there a magnitude effect? *Journal of the Experimental Analysis of Behavior* 81: 39–50.
Herrnstein, R. J. 1990. Rational choice theory: necessary but not sufficient. *American Psychologist* 45: 356–367.
Hilgard, E. R. and Marquis, D. G. 1940. *Conditioning and Learning*. New York: Appleton-Century.
Huizinga, J. 1924. *The Waning of the Middle Ages*. New York: St. Martin.
James, W. 1890. *Principles of Psychology*. New York: Holt.
Johnson, A. and Redish, A. D. 2007. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience* 12: 483–488.
Kehoe, E. J., Graham-Clark, P., and Schreurs, B. G. 1989. Temporal patterns of the rabbit's nictitating membrane response to compound and component stimuli under mixed CS-US intervals. *Behavioral Neuroscience* 103: 283–295.
Kirby, K. N. 2006. The present values of delayed rewards are approximately additive. *Behavioural Processes* 72: 273–282.
Kirby, K. N. and Guastello, B. 2001. Making choices in anticipation of similar future choices can increase self-control. *Journal of Experimental Psychology: Applied* 7: 154–164.
Klein, B. and Leffler, K. B. 1981. The role of market forces in assuring contractual performance. *Journal of Political Economy* 89: 615–640.
Kuhn, M., Lutolf, M., and Reinhart, W. H. 1999. Ondine's Curse. *Respiration International Review of Thoracic Disease* 663: 265.
Laibson, D. 2001. A cue-theory of consumption. *Quarterly Journal of Economics* 66: 81–120.
Lea, S. E. G. and Webley, P. 2006. Money as tool, money as drug: The biological psychology of a strong incentive. *Behavioral and Brain Sciences* 29: 161–209.
Lerner, J. S. and Tiedens, L. Z. 2006. Portrait of the angry decision maker: How appraisal tendencies shape anger's influence on cognition. *Journal of Behavioral Decision Making* 19: 115–137.
Loewenstein, G. F. 1999. A visceral account of addiction. In *Getting Hooked: Rationality and Addiction*, eds. J. Elster and O. J. Skog. Cambridge: Cambridge University Press: 65–92.
Maisto, S., Lauerman, R., and Adesso, V. 1977. A comparison of two experimental studies of the role of cognitive factors in alcoholics drinking. *Journal of Studies on Alcohol* 38: 145–49.
Marks, I. 1997. Behavior therapy for obsessive-compulsive disorder: A decade of progress. *Canadian Journal of Psychiatry* 42: 1021–1027.
Mazur, J. E. 2001. Hyperbolic value addition and general models of animal choice. *Psychological Review* 108: 96–112.
Monterosso, J., Ainslie, G., Toppi-Mullen, P., and Gault, B. 2002. The fragility of cooperation: A false feedback study of a sequential iterated prisoner's dilemma. *Journal of Economic Psychology* 234: 437–448.
Pap, A. 1961. Determinism, freedom, moral responsibility, and causal talk. In *Determinism and Freedom in the Age of Modern Science*, ed, S. Hook. Collier: New York.
Peciña, S., Smith, K. S., and Berridge, K. C. 2006. Hedonic hot spots in the brain. *The Neuroscientist* 12: 500–511.

Prelec, D. and Bodner, R. 2003. Self-signaling and self-control. In *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, eds. G. Loewenstein, D. Read, and R. Baumeister, 277–298. New York: Russell Sage.

Rolls, E. T. 2005. *Emotion Explained*. Oxford: Oxford University Press.

Rozin, P., Markwith, M., and Stoess, C. 1997. Moralization and becoming a vegetarian: The transformation of preferences into values and the recruitment of disgust. *Psychological Science* 8: 67–73.

Russell, J. M. 1978. Saying, feeling, and self-deception. *Behaviorism* 6: 27–43.

Savastano, H. I., Hua, U., Barnet, R. C., and Miller, R. R. 1998. Temporal coding in Pavlovian conditioning: Hall-Pearce negative transfer. *Quarterly Journal of Experimental Psychology* 51: 139–153.

Schelling, T. C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schelling, T. C. 1986. The mind as a consuming organ. In *The Multiple Self*, ed. J. Elster, 177–195. Cambridge: Cambridge University Press.

Shizgal, P. and Conover, K. 1996. On the neural computation of utility. *Current Directions in Psychological Science* 5: 37–43.

Skinner, B. F. 1948. Superstition in the pigeon. *Journal of Experimental Psychology* 38: 168–172.

Sternbach, R. A. 1968. *Pain: A Psychophysiological Analysis*. New York: Academic.

Sully, J. 1884. *Outlines of Psychology*. New York: Appleton.

Tolman, E. C. 1939. Prediction of vicarious trial and error by means of the schematic sowbug. *Psychological Review* 46: 318–336.

Weber, M. 1904/1958. *The Protestant Ethic and the Spirit of Capitalism*. New York: Charles Scribners.

Wegner, D. M. 2002. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Chapter 8
From Belief Revision to Preference Change

Till Grüne-Yanoff and Sven Ove Hansson

Abstract We propose to model the consistency-preserving aspect of preference change after the fashion of belief revision. First, we discuss the formal properties of the preference notion. Second, we discuss the various consistency requirements imposed on preference sets. Third, we discuss representations of consistency-driven preference change and compare them to models of belief change. Last, we discuss the specific needs of introducing a priority index in models of preference change. We conclude that while the general input-assimilating framework from belief change can be transferred to preference change, several modifications are necessary. In particular, the input model has to be complicated with the introduction of a distinction between primary (non-linguistic) and secondary (linguistic) inputs. Sentential representation has to be used with somewhat more caution for preferences than for beliefs. The priority-setting mechanism has to be adjusted, and priority-related information must be included in the inputs.

8.1 Introduction

How should a formal theory of preference change be constructed? In order to get a systematic grip on that issue, we have chosen to attack it from two sides. First and most obviously, we draw on previous studies of preferences, both preference logic and more informal discussions on preferences in the social sciences. Secondly, we compare preference change to belief change. There are important similarities, but – as we will soon see – also major differences between these two areas of formalized philosophy. Contrary to preference change, belief change is an established field. In formal epistemology and theoretical computer science, a large variety of formal models of belief change have been developed, for both descriptive and normative

T. Grüne-Yanoff
Helsinki Collegium of Advanced Studies and Royal Institute of Technology, Stockholm
e-mail: till.grune@helsinki.fi

S.O. Hansson (✉)
Royal Institute of Technology, Stockholm
e-mail: soh@kth.se

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

purposes (Gärdenfors 1988; Hansson 1999). In this contribution we intend to identify the central aspects of standard belief change models, discuss their applicability to preference change, and in this way put focus on several important issues in preference change. As we will show, some of the central features of belief change models can be used in preference change, but a number of extra features are also required that distinguish preference change models from the main models of belief revision.

Mental changes, or changes in mind, can take many forms, of which changes in beliefs and preferences are only two. Our norms, our emotions, our patterns of argumentation, our ideologies, etc. are also subject to change. Preference changes should not be seen as independent of these other types of change. Some of the most important questions that we need to clarify concern the interconnections between changes in these various compartments of the mind. How are for instance preferences affected by changes in beliefs, and the other way around? In order to investigate such issues we need models that represent larger parts of a state of mind than its preferences.

It is probably a good strategy to develop workable models of preference change before we try to develop such larger models. Discussing the makeup of such a simple model will be the main aim of this paper. But it is also advisable to construct a model of preference change such that it is embeddable into larger models of changes in mental states. This is a factor that we will pay particular attention to in what follows.

In Section 8.2 we investigate how preferences relate to four other conceptual categories that are important in the dynamics of mind and action, namely values, norms, choices, and beliefs. In Section 8.3 we discuss how preferences and their relata should be represented, and in Section 8.4 we investigate what general rationality constraints should apply to all preference states. Section 8.5 is devoted to fundamental issues in the representation of change, such as the typology of change operations and the role of inputs in these operations. In Section 8.6 we discuss what mechanisms should be used to select among alternative outcomes that all satisfy the rationality constraints. Section 8.7 concludes.

8.2 Preferences, Values, Norms, Choices, and Beliefs

First of all, the category of preferences has to be positioned among the various other categories that may be subject to formal analysis. We have chosen to focus on four such categories that seem to be particularly relevant for the comprehensive formal characterization of preferences, namely values, norms, choices, and beliefs.

8.2.1 Values

Preferences are expressions of values. From a structural point of view, the value concepts that we use in ordinary language as well as in more specialized discourse can be divided into two major categories. The *monadic* (classificatory) value concepts,

such as 'good', 'very bad', and 'worst', evaluate a single object. The *dyadic* (comparative) value concepts, such as 'better', 'worse', 'at least as good as', and 'equal in value to' make a value-based comparison between two objects. Preferences, as they are usually conceived, have their place in this category. To say that someone prefers A to B is synonymous with saying that according to some of that person's values, A is better than B.[1]

The common dyadic value concepts are usually taken to be interdefinable, hence it is assumed that A is better than B if and only if it is both the case that A is at least as good as B and that A is not equal in value to B. The subject-matter of preference logic is usually taken to cover all the dyadic value concepts. This will be our approach here. In other words, we will consider the topic of preference change to cover not only changes in what the subject prefers to something else but also in what the subject considers to be of equal value as something else.

There are close structural relationships between monadic and dyadic values. It would seem paradoxical to claim both that A is better than B and that B is best. It would be almost equally strange to claim both that A is better than B and that A is bad whereas B is good. It is generally accepted in formal studies of preferences that the monadic value predicate 'best' can be defined in terms of dyadic value predicates. An object A is best among a group of objects if it is better than all other objects in that group (or alternatively if no object in that group is better than A). At the other end of the value-scale, 'worst' is defined analogously (Hansson 2001a, pp. 115–116). Proposals are also available for the definition of 'good' and 'bad' in dyadic terms. According to one such proposal, a value-object is good if and only if it is better than its negation, and it is bad if and only if it is worse than its negation (Brogan 1919). According to the other major proposal, a value-object is good if and only if it is better than some proposition that is neither better nor worse than its negation. Similarly, it is bad if and only if it is worse than some proposition that is neither better nor worse than its negation (Chisholm and Sossa 1966). Both these definitions have the disadvantage of only being applicable to negatable value objects.

If monadic values can be defined in terms of dyadic values, then an account of preference change, i.e. change in dyadic values, will generate a derivative account of changes in monadic values, so that the logic of preference change becomes a general logic of value change.

8.2.2 Norms

Even among philosophers otherwise committed to fine linguistic distinctions, the distinction between norms and values is often overlooked. There is in fact an

[1] For some technical purposes, value predicates with more than two referents may be useful, such as the four-termed "x is preferred to y more than z is preferred to w" (Packard 1987). Our focus here will be on the dyadic concepts.

essential difference: norms are directly action-guiding whereas values are not. Hence, suppose that we have a choice between three exhaustive and mutually exclusive action alternatives A, B, and C. The statement that A is better than each of B and C, and B better than C, is unproblematically compatible with each of the following three statements (1) A is obligatory whereas B and C are forbidden, (2) A and B are both permitted whereas C is forbidden, and (3) A, B, and C are all permitted. More generally speaking, even if we know what values someone assigns to a set of alternatives, we cannot infer from this what normative statements she endorses. The best alternative may be supererogatory (i.e. good, but not obligatory; cf. Chisholm 1963), or the normative appraisal may be based on a satisficing account of normativity (Slote 1984).

Of course, if we adopt the principle that maximal value-production is required of all agents, then the normative status of an action can be derived from its value status. However, this principle is implausible, since it rules out supererogatory acts and, even more importantly, limits the freedom of the agent to a choice among a few value-maximal alternatives (Hansson 2006).

Although norms are not in general derivable from preferences (or from other expressions of value), norms and values are not fully independent of each other. It would for instance not seem credible to claim that the action A is better than all alternative actions and at the same time maintain that A is forbidden whereas all its alternatives are permitted.[2] Criteria of coherence can be applied to combinations of norms and preferences, but these criteria cannot be assumed to be sufficiently specified to make norms derivable from preferences.

8.2.3 Choices

There is a strong tradition, particularly in economics, to equate preference with choice. Preference is considered to be hypothetical choice, and choice to be revealed preference. Hence, the Arrovian framework in social choice theory "conflates 'choice' and 'preference"', and treats these "as essentially synonymous concepts.... A preference is a potential choice, whereas a choice is an actualized preference" (Reynolds and Paris 1979). Arrow himself has defined preference as "choice from two-member sets" (Arrow 1977, p. 220). The same approach dominates in other areas of economics.

However, the conflation of choices and preferences is a rather far-stretched idealization that is not adequate for all purposes. In fact, choices and preferences are entities of quite different categories. Preferences are parts of *states of mind*. That a person prefers A to B means that she considers A to be better than B. Choices are

[2] This intuition is supported by a plausible deontic principle, namely contranegativity (OX & ($\neg X \geq \neg Y$) $\rightarrow$ OY, where O is the ought operator). To see this, let B be any of the alternatives to A. Then according to the assumptions of the example, $A > B$ and $O \neg A$. Contranegativity yields $O \neg B$ so that $\neg O \neg B$ does not hold, i.e. B is not permitted.

actions. That someone has chosen A means that she has actually selected A (irrespectively of whether she judges A to be better than its alternatives). Even in market behaviour, the primary subject-matter of economic theory, there are several types of situations in which choice and preference clearly do not coincide (Sen 1973). In particular, in markets and elsewhere, a person can select from alternatives that she is indifferent between or considers to be incomparable (Ullmann-Margalit and Morgenbesser 1977). It is also possible to have preferences over alternatives that one cannot choose between. Suppose that there are two prizes in a lottery: a luxury cruise worth € 10,000 and an account of € 10,000 at your local grocery store that you can use to buy food in the years to come. You may then very well prefer winning the luxury cruise to winning the account at the grocery, but since you cannot choose what to win – if you could it would not be a lottery – this preference is not directly connected to choice. (As this example may illustrate, you may prefer *winning A* to *winning B* even though you would, in a direct choice between A and B, choose B rather than A.)

Hence, although preferences and choices are often conflated, they are very different in nature and should be treated as phenomena on different levels in a model of mind and action. Preferences influence choices, just as beliefs do, but they should be carefully kept apart from choices in a formal model of mental changes.

8.2.4 Beliefs

Both beliefs and preferences are parts of a person's state of mind. Beliefs refer to the realm of facts and preferences to the realm of value. We have learned to keep these realms apart, and of course the distinction should not be blurred. However, this does not mean that preferences and beliefs are independent of each other, so that any combination of preferences is compatible with any combination of beliefs. Instead, their relation is similar to that between values and norms: neither is derivable from the other but a change in one of the two categories can have impacts on the other.

The influence of beliefs on preferences is largely uncontroversial since it is generally accepted that preferences should be factually well-informed. New information often leads us to modify our preferences. Some preferences would be considered irrational if we have certain beliefs. Hence, if a person believes (correctly) that Le Corbusier and Charles-Edouard Jeanneret-Gris were one and the same person, then it would be incoherent of her to prefer houses built by Le Corbusier to houses built by Charles-Edouard Jeanneret-Gris.

The reverse influence, from preferences to beliefs, is more controversial. Wishful thinking, i.e. believing that things are as we wish them to be, often leads us astray (Hansson 2006). However, other, more sophisticated influences of preferences on beliefs seem to be justified. In particular, our values can influence the standards of evidence, or burdens of proof, that we assign to different potential beliefs. Hence, our strong preference for safety in a medical context makes us put high demands on the evidence before we allow ourselves to believe that a new drug is safe for

humans. In comparison, we tend to require less stringent evidence before we come to believe that a drug has a serious side-effect. Although this asymmetry in standards of evidence is not uncontroversial, it should not be excluded in a general framework for studies of mental change.

8.2.5 Summary

We can summarize these considerations as follows: By changes in "preferences" we actually mean changes in value comparisons, i.e. in those values that we express with dyadic value statements. This also includes for instance a change from regarding two objects as incomparable to considering them to be of equal value. Preferences (in this wide sense) are closely related with monadic value concepts, and at least some monadic value statements are derivable from preferences.

Preferences, norms, and beliefs are all parts of the mental state, or state of the mind. These three categories are distinctly different, and a statement belonging to one of them cannot be synonymous with a statement belonging to one of the others. Nevertheless there are relationships among the three categories. Even if all beliefs in a state of the mind are internally coherent, and the same applies to the preferences and the norms, the combination of the three components may be incoherent.

Finally, choices are not parts of the state of the mind. In a model representing both actions and the mind, choices should be represented among the actions and preferences as parts of the state of mind. Just like beliefs and norms, preferences can influence choices. Yet choices and preferences are not identical.

8.3 The Representation of Preferences

The choice of a preference representation model has direct influence on the framework for preference change. We propose a representation of preferences as binary relations over sentences. This distinguishes our framework from expected utility functions over goods bundles on the one hand (as standardly used in microeconomics) and from modal logic frameworks on the other (for examples, see van Benthem, Chapter 3, this volume). We justify our choice by simplicity and convenience. It is convenient, because a representation based on sentences immediately places our framework close to standard models of belief revision, thus allowing easy comparison. It is further convenient because a large part of the social science literature naturally relates to this formal framework. It is simple because it leaves out probabilistic information. Deliberately forgoing this extra information forces us to model preference change without recourse to these parameters. Of course, this limits the application of our framework to many real-world situations, in which probability change often plays an important role (for a model including probabilistic change, see Bradley, Chapter 11, this volume). Yet it should also be pointed out that at the basis

of any expected utility theory lies a simple preference ordering of the sort modelled here. Therefore our framework can be seen as a necessary basis of expected utility theory, and thus as compatible with it.

8.3.1 Relata

In order to represent preferences we need a representation of that which they refer to, the relata. A manageable formal representation of the relata will require some degree of simplification, since in non-regimented language, all sorts of abstract and concrete entities can serve as the relata of preference relations. Thus, one may prefer coffee to tea, logic to postmodern literature theory, or novels to poetry. In spite of this, preference theory has been almost exclusively restricted to two representations of relata: Either relata are taken as primitives, or they are taken to be sentences representing states of affairs.

The use of states of affairs to represent relata can be defended with reference to the ease with which we can translate talk about other types of relata into talk about states of affairs. This translation has often been taken as unproblematic. Hence R. Lee (1984, pp. 129–130) claimed that "all preferences can be understood in terms of preference among states of affairs or possible circumstances. A preference for bourbon, for example, may be a general preference that one drink bourbon instead of drinking scotch" (Cf. von Wright 1963, p. 12; Trapp 1985, p. 303).

Arguably, it is not quite as simple as that. A person's preference for one musical piece over another, for example, cannot be translated into a single preference for one state of affairs over another. Instead, it can be represented by a conglomeration of preferences referring to these pieces of music: she may prefer states of affairs in which she *plays* the first rather than the second piece, but she may also prefer a state of affairs in which she *listens* to the first rather than the second, etc. To dissolve this ambiguity, one needs to investigate the context in which a preference over the primitives was expressed.

In spite of not being a perfect representation, sentences expressing states of affairs are the best general-purpose representation of the relata of preferences. This is a welcome conclusion, since sentences are also the best general-purpose representation of beliefs, and the best general-purpose representation of the objects of norms. If we are interested in building general models of mental states that include several of these elements, then the use of sentential representation is highly advisable since it provides unity and thereby simplifies investigations of connections between the categories.

8.3.2 The Comparative Predicates

There are two fundamental dyadic (comparative) value concepts, namely 'better' (strict preference) and 'equal in value to' (indifference) (Halldén 1957, p. 10).

We will use the symbols $>$ and $\equiv$, respectively, to denote them.[3] Furthermore, in accordance with a long-standing philosophical tradition, we will take $A > B$ to represent "B is worse than A" as well as "A is better than B", thus abstracting from whatever psychological or linguistic asymmetries that may persist between betterness and worseness.

The following two properties of the two comparative relations will be taken to be part of the meaning of the concepts of (strict) preference and of indifference:

$$A > B,\ B > A,\ \text{and } A \equiv B \text{ are pairwise mutually exclusive} \tag{8.1}$$

$$\text{If } A \equiv B \text{ then } B \equiv A \tag{8.2}$$

These two conditions combine to ensure that $>$ and $\equiv$ give rise to a fourfold classification of all pairs of objects of comparison:

$$A \text{ is equal in value to } B (A \equiv B) \tag{8.3}$$

$$A \text{ is strictly preferred to } B\ (A > B) \tag{8.4}$$

$$B \text{ is strictly preferred to } A\ (B > A) \tag{8.5}$$

$$\text{The comparison between } A \text{ and } B \text{ is undetermined } (A <> B)^4 \tag{8.6}$$

We will also assume that indifference is reflexive:

$$A \equiv A \tag{8.7}$$

Preference logic is usually not performed with $>$ and $\equiv$ as primitives. Instead, it is common to use 'at least as good as' (or more precisely: 'better than or equal in value to'), denoted $\geq$, as the sole primitive. With our three basic assumptions, the two sets of primitives are interdefinable with the definition $A \geq B \leftrightarrow (A > B) \vee (A \equiv B)$ in one direction and the two definitions $A > B \leftrightarrow (A \geq B) \,\&\, \neg(B \geq A)$ and $A \equiv B \leftrightarrow (A \geq B) \,\&\, (B \geq A)$ in the other (Hansson 2001a, p. 19).

The choice of primitives (either $\geq$ alone or both $>$ and $\equiv$) is a choice between formal simplicity ($\geq$) and conceptual clarity ($>$ and $\equiv$) (Hansson 2001b, pp. 321–322). For most purposes, the choice is not important, but for some basic conceptual purposes it is necessary to choose the option with $>$ and $\equiv$. (This will be exemplified in Section 8.5.3.) We therefore propose that $>$ and $\equiv$ be used as the primitive comparative value terms.

---

[3] For simplicity, we will leave out from explicit discussion two important topics in a formal account of preferences and related concepts: (1) The set of alternatives or options that $>$ and $\equiv$ range over. (2) The standard of evaluation, such as moral value, aesthetic value, or value *tout court*, that they refer to.

[4] The fourth category consists in the absence of any of the other three. This has consequences for its representation. Thus whereas $A > B$ will hold in a preference set $\mathbf{S}$ if and only if $A > B \in \mathbf{S}$, and similarly $A \equiv B$ holds in $\mathbf{S}$ if and only if $A \equiv B \in \mathbf{S}$, $A <> B$ holds in $\mathbf{S}$ if and only if $\mathbf{S} \cap \{A > B,\ B > A,\ A \equiv B\} = \emptyset$.

8.3.3 Preference States

As we have already emphasized, preference states are excised parts of the mental state, or state of the mind. A preference state cannot exist in isolation; therefore when we choose to treat it as a self-sufficient entity this is an idealization.

The most obvious way to represent a preference state is probably to let it be represented by the set of all sentences in the preference language that it endorses. This construction is similar to that of a belief set (corpus) that consists of all the sentences that the subject believes in, or is committed to believe in. In analogy to belief sets, a set consisting of all the sentences (in a given language) that represent preferences held by the subject will be called a *preference set*.

Just like belief sets, preference sets as defined here are closed under logical consequence. Hence, a person who subscribes to the preference sentence $A \geq B$ is assumed to also subscribe to the disjunctive sentence $(A \geq B) \vee (A \geq C)$. Furthermore, if transitivity is one of the background conditions, and she subscribes to both $A \geq B$ and $B \geq C$, then we assume that she also subscribes to $A \geq C$. The logical closure of the preference set may be seen as the outcome of a reflective equilibrium. Somewhat more modestly, we may interpret a preference set as representing, not the set of actually endorsed preference sentences, but the set of preference sentences that the agent is committed to endorse.[5]

This construction has the advantage of conforming with representations that we have reasons to choose for other parts of the mental state. If we have both a belief set and a preference set, then they can be combined in ways that facilitate the formal treatment of connections between the two.

However, the logical closure of belief sets and preference sets also has drawbacks. Many distinctions are lost in the process of logical closure. This problem has been highlighted in previous studies of belief change (Hansson 1999, pp. 17–24). One major problem with belief set models is that they allow for only one inconsistent state. The reason for this is that there is only one set that is both inconsistent and logically closed (namely the whole language). This is an unsatisfactory property of belief set models, since intuitively speaking there are many ways to hold inconsistent beliefs. This is a problem that belief modelling has in common with preference modelling. For example, it may be important for the understanding of possible preference state changes whether the agent violated transitivity (where consistency could for instance be restored by removing any one of the three relations $A > B$, $B > C$, $C > A$), or whether she violated asymmetry (where consistency could be restored by removing either $A > B$ or $B > A$).

One possible remedy is to replace belief sets by belief bases, sets that are not closed under logical consequence. Hence, instead of the belief set $\mathrm{Cn}(\{A, B\})$ we can use one of the belief bases $\{A, B\}$, $\{A\ A \rightarrow B\}$, $\{A \vee B,\ A \leftrightarrow B\}$ etc., all of which have the same logical closure and therefore correspond to the same belief set. For each belief set there are many belief bases. To the extent that we can give the distinction between belief bases a meaningful interpretation, much more

[5] This distinction was introduced for belief sets by Isaac Levi (1974, 1977, 1991).

information can be conveyed in this representation. Arguably there is a meaningful such representation, namely that the belief base consists of those beliefs that have an independent justification.

The same argumentation applies to preference states. In the same way, we can replace the preference set by a preference base that contains those preferences that the agent has actively accepted and that have survived subsequent changes. Elements of the preference base thus contrast with the merely derived preference statements that form the rest of the preference set. This allows us to distinguish between different ways of holding inconsistent preferences. For example, if someone prefers drinking tap water to drinking mineral water ($T > M$) then it follows that she either prefers drinking tap water to drinking mineral water or prefers drinking sewage water to drinking tap water, $(T > M) \vee (S > T)$. However, this latter element in her state of preferences is merely derived and will disappear as soon as $T > M$, from which it was derived, disappears. If she adopts the new preference $M > T$, to replace $T > M$, the option of retaining $(T > M) \vee (S > T)$ (which with $M > T$ yields $S > T$) does not even arise. In contrast, in a framework with preference sets, $(T > M) \vee (S > T)$ does not disappear automatically when $T > M$ is given up. Its elimination will have to be ensured with some priority-setting mechanism (see further Section 8.6.3).

8.3.4 Summary

In accordance with tradition, we propose the use of sentences denoting states of affairs as a general representation of the relata of value comparisons. This choice will facilitate combinations of preference states with other parts of mental states, since such sentences are also the best general-purpose representations of beliefs and of the objects of norms. We propose the use of strict preference ($>$) and indifference ($\equiv$) as the primitive comparative predicates, since they are conceptually more fundamental than the alternative primitive predicate ($\geq$).

There are two major alternative ways to combine these elements into preference states, namely (logically closed) preference sets and (logically open) preference bases. Both types of models are worth further investigations. In combined models including other mental entities such as beliefs it will mostly be advisable to use either closed or open representation throughout.

8.4 Integrity Constraints

8.4.1 Integrity Constraints Versus Priorities

With integrity constraints we mean requirements that a preference state has to satisfy in order to be an adequate representation of preferences according to the standards of the chosen model. Integrity constraints are exceptionless, i.e. they apply to all

preference states that are the outcome of some operation of change. (We can leave it open whether such an operation can be applied to a preference state not satisfying the constraints; the essential criterion is that the posterior states satisfy them.) Typical examples of integrity constraints are logical closure in preference set models and transitive closure in models of rational preferences that require transitivity.

Integrity constrains should be distinguished from input constraints that come with the specific input, and therefore do not apply to all preference states. A typical example of an input constraint is the requirement that the outcome of contracting some preference state by some non-tautological sentence (such as $A > B$) should be a new preference state not containing or endorsing that sentence.

Integrity constraints should also be distinguished from priorities and priority-setting mechanisms. A priority-setting mechanism, such as a selection function, incision function, or entrenchment relation, tells us for instance which of two elements in a preference set we should retain when the combination of integrity and input constraints prevents us from retaining both of them.

We propose, as a general strategy, that if there is a choice between expressing a condition as an integrity constraint or as a priority-setting principle, then the former option should be chosen. A major reason for this is that many integrity constraints can be included in the logic, which allows for a more unified formal treatment. In the next section we will proceed to show how this is done. Integrity constraints can often perform the function of priority-setting criteria, e.g. they can contribute to determining the choice between alternative ways to restore consistency when a new preference is added that is consistent with the set of previous preferences. In this way, priorities can to some extent be *endogenised* through incorporation into the logic (see further Section 8.6.3).

8.4.2 Formalizing Integrity Constraints

Many integrity constraints take the form of rationality postulates such as transitivity, $(X > Y)\ \&\ (Y > Z) \rightarrow (X > Z)$. In order to see how such postulates can be incorporated into the logic it is useful to focus on the consequence operator that is associated with the logic. The minimal consequence operator for our purposes is the classical truth-functional consequence operator Cn_0, such that for each set S of sentences, $\mathrm{Cn}_0(S)$ is the set of its truth-functional consequences. Rationality postulates will be represented by stronger consequence operators. Let T be a set of preference postulates. Then Cn_T is the operator such that $\mathrm{Cn}_T(S)$ consists of the logical consequences that can be obtained from S, using both truth-functional logic and the postulates in T. In other words,

$$\mathrm{Cn}_T(S) = \mathrm{Cn}_0(s(T) \cup S), \tag{8.8}$$

where $s(T)$ is the set of substitution instances of elements of T. As an example, if $(X \geq Y) \vee (Y \geq X) \in T$ and $\neg(A \geq B) \in S$, then $B \geq A \in \mathrm{Cn}_T(S)$.

(Clearly, $(A \geq B) \vee (B \geq A) \in s(T)$, and the rest follows truth-functionally.) The important observation that makes this construction work is that whenever Cn_0 is Tarskian consequence operator then so is Cn_T (Hansson 2001a, pp. 35–36).

For any model $\mathcal{M}$ of a preference state, let $|\mathcal{M}|$ be the set of preference sentences that it endorses. Then we can define $\mathcal{M}$ as consistent if and only if $\mathrm{Cn}_T(|\mathcal{M}|)$ is consistent. In this way, integrity constraints (such as transitivity) are expressible in terms of logical consistency. This makes the formal treatment much more unified than if we had to treat each integrity constraint separately.

8.4.3 Internal Integrity Constraints

We can divide the integrity constraints concerning preferences into two major categories, namely those that refer to relations among preferences and those that refer to relations between preferences and other mental objects. Of course only the former are directly relevant in a model that contains only preferences and no other mental entities.

One major class of such internal integrity constraints are properties such as completeness, acyclicity, transitivity, IP-, PI-, II-, and PP-transitivity that facilitate the use of the preference state for action-guiding purposes. The decision which of these potential constraints to include in a preference model will depend largely on the purpose of the model. In a descriptive model most of these properties will probably not be satisfied for the simple reason that people tend to violate them. Two major potential reasons have been given why one should honour these constraints. First, the standard *meaning* of preferences is held to be partly constituted by these constraints (Davidson 1980, p. 273). Secondly, it may be argued that preferences should have such a structure that they can be used to guide our choice among the alternatives that they cover. To make consistent choice-guidance possible, some integrity constraints will have to be satisfied.[6]

There is also another class of constraints, namely those that refer to logical relations among relata. Such relations are often excluded by the simplifying assumption that all objects of preferences are mutually exclusive, i.e. none of them is compatible with, or included in, any of the others. However, actual agents often have preferences that refer to compatible relata. One can prefer ice cream to fruit cake although it is quite feasible to have both. We should expect there to be logical connections among preferences that refer to logically related relata.

The following are two plausible such integrity constraints:

$$(X \geq Y) \rightarrow (X \geq (X \vee Y)) \;\&\; ((X \vee Y) \geq Y) \; (\textit{disjunctive interpolation}) \quad (8.9)$$

and its weaker variant

$$(X \equiv Y) \rightarrow (X \equiv (X \vee Y)) \quad (8.10)$$

[6] See Hansson (2001a, pp. 23–26) for a detailed discussion of what integrity constraints a preference relation has to satisfy in order to be useful for action-guidance.

Logicians have tried to construct these connections in two ways. The most common of these is the *holistic* approach that takes preferences over wholes for basic and uses them to derive the other preferences. The wholes chosen for this purpose have usually been possible worlds. Preferences over sentences can then be derived in various ways from preferences over the worlds in which these sentences are true (Hansson 2001a, pp. 57–113). Unfortunately this construction has the disadvantage of blatant cognitive unrealism. In practice, we are not capable of deliberating on anything approaching the size of completely determinate possible worlds.[7] The use of smaller holistic objects ("myopic holism") has been proposed as a means to overcome these difficulties (Hansson 2001a, p. 59).

The other approach has been called *aggregative*. It takes small units to be the fundamental bearers of value, and the values of larger entities are obtained by aggregating the values of these units. This means that preferences over states of affairs that describe only a small part of the state of the world are the fundamental bearers of value, and preferences over truth-functional combinations of these states are derived from them. A precise numerical aggregative model was developed by Warren Quinn on the basis of a proposal by Gilbert Harman. In that model (intrinsic) values are assigned to certain basic propositions, and precise rules are given for deriving the values of truth-functional combinations of these basic units (Harman 1967; Quinn 1974; Oldfield 1977; Carlson 1997; Danielsson 1997). This construction is based on the assumption that there are completely separable and evaluatively independent bearers of value, and that a numerical representation is available in which aggregate value is obtainable through addition of the values of these isolable units (Spohn 1978, pp. 122–129). Needless to say, these are strong and implausible assumptions (Moore 1903, p. 28).

We will leave open the issue how the relations between preference statements with logically related relata should be constructed. It should at any rate be clear that such connections belong among the integrity constraints in a model of preference change.

8.4.4 External Integrity Constraints

In models containing other mental elements in addition to preferences, integrity constraints should be included that refer to these combinations. Without going into details, we would like to mention a few examples.

[7] In studies of concepts or phenomena that one considers to be independent of cognition, it may be reasonable to abstract from cognitive limitations and use models with completely determinate possible worlds. This applies to some concepts of possibility; R.M. Adams has indeed claimed that "possibility is holistic rather than atomistic, in the sense that what is possible is possible only as part of a possible completely determinate world" (Adams 1974, p. 225). However, this argument for possible world modelling is not applicable to evaluative and normative concepts.

One plausible such constraint concerning preferences and *norms* is the following, for the ought operator O:

$$OX \,\&\, (\neg X \geq \neg Y) \rightarrow OY \text{ (}\textit{contranegativity}\text{)} \tag{8.11}$$

If preferences are complete, this constraint will be equivalent with $OX \,\&\, \neg OY \rightarrow (\neg Y > \neg X)$. Hence, if you ought to pay your debt to your destitute neighbour, but you are not obliged to pay your debt to the car-dealer, then it would be worse of you not to pay your neighbour than not to pay your car-dealer.

Concerning the relationship between preferences and *beliefs*, one obvious and fairly plausible principle is *intersubstitutivity of relata believed to be logically equivalent*. In other words, if the agent believes two relata to be logically equivalent, then they should be exchangeable with no impact on truth or endorsement. For example, if a persons prefers ingesting 10 g of Vitamin C a day to not doing so, and believes that Vitamin C is ascorbic acid, then she is also committed to prefer ingesting 10 g of ascorbic acid a day to not doing so.

The topic of integrity constraints that connect preferences and beliefs is closely connected with central issues both in epistemology and value theory. As one example, we may ask: How and under what conditions should a change in factual beliefs give rise to a change in the values held by the subject? That such changes take place is obvious. Suppose that you prefer spending the evening with one person rather than another. Unless these are very entrenched preferences, they can be reversed if you acquire relevant new information about one of the persons. A common reaction to examples like this is that such a preference change is superficial or perhaps even illusory, and that your underlying preferences for what kind of person you spend the evening with are unchanged (cf. Stigler and Becker 1977). But can these underlying preferences be clearly delineated and in either case, what are the implications for preference change? These are issues worth a careful investigation.

8.4.5 Summary

By integrity constraints we mean requirements to be satisfied by all preference states, or at least by all preference states that are arrived at by an operation of change. Integrity constraints have the major advantage of being incorporable into the logic, which allows for a unified formal treatment. They should be distinguished from (1) input constraints that apply only to preference states that result from a specific operation of change, and (2) priority criteria that instruct us on the choice between different ways to satisfy the integrity and input constraints.

A model of preference change should include integrity constraints that mirror the logic of preferences. In addition, a model that includes other mental entities than preferences should include connecting constraints, such as constraints connecting preferences to beliefs and norms.

8.5 The Representation of Change

8.5.1 Input-Assimilation

The major models of belief change are *input-assimilating*: the belief state (either a belief set or a belief base) is exposed to an input, which imposes input constraints. As a result, the belief state is changed. The outcome of these changes is determined by the combination of integrity constraints, input constraints, and priority criteria, as outlined above.

We will adopt this general approach, but we will be very open concerning the nature of the inputs.

8.5.2 The Types of Belief Change

As a starting-point, let us consider the types of change that have been developed in belief change theory. There are three dominating kinds of belief change. In *expansion*, a specified sentence is added to the belief set. In *revision*, a specified sentence is added, and if needed, other sentences are removed in order to retain consistency. In *contraction*, a specified sentence is removed.

In addition to these, several other types of belief change have been proposed:

Consolidation: A belief base is made consistent by removing enough of its more dispensable elements. (Hansson 1994)[8]

Semi-revision: An input sentence that contradicts previous beliefs is accepted if it has more epistemic value than the original beliefs that contradict it. Otherwise the original belief state is retained. (Hansson 1997; Olsson 1997)

Selective revision: This is a generalization of semi-revision in which it is possible for only a part of the input information to be accepted. (Fermé and Hansson 1999; Gabbay 1999)

Shielded contraction: This is a version of contraction in which some non-tautological beliefs cannot be given up; they are shielded from contraction. (Makinson 1997; Fermé and Hansson 2001)

Replacement: One sentence is replaced by another in a belief set. Hence $K\left|_{q}^{p}\right.$ is a belief set that contains q but not p. Replacement can be used as a "Sheffer stroke" for belief revision. Contraction by p can be defined as $K\left|_{\mathrm{T}}^{p}\right.$, revision by p as $K\left|_{p}^{\perp}\right.$, and expansion by p as $K\left|_{p}^{\mathrm{T}}\right.$. where $\perp$ is falsum and T is tautology. (Hansson 2009)

Multiple contraction and revision: The simultaneous contraction (revision) of more than one sentence. (Fuhrmann and Hansson 1994)

[8] This operation cannot be meaningfully applied to belief sets, since there is only one inconsistent belief set. Once inconsistency has been reached in a belief set system, all distinctions have been lost, and they cannot be regained in an operation of consolidation.

An important feature in all but one of these operations is that they are defined in terms of one or several input sentences, although they differ in what is done with that input sentence (remove it, add it, add it if it has a sufficiently high position in the priority ordering, etc.). The exception is of course consolidation.

8.5.3 Three Basic Types of Preference Change

We will not take it for granted that the inputs of preference change should be determined by sentences in the same way as in belief change. Therefore, we will begin by classifying preference changes in terms of the relationships between the prior and the posterior preference state, for the moment making no assumption about the nature of the input.

Given two relata A and B, there are three fully specific comparative statements that can be made about them: $A > B$, $B > A$ and $A \equiv B$. There are also some other, less specified types of statements that we can make. Three of these are practically relevant, namely

$$A \geq B, \text{ that is equivalent to } (A > B) \vee (A \equiv B), \tag{8.12}$$

$$B \geq A, \text{ that is equivalent to } (B > A) \vee (A \equiv B), \text{ and} \tag{8.13}$$

$$A <> B \text{ that holds if and only if neither } A > B, B > A, \text{ nor } A \equiv B \text{ holds.} \tag{8.14}$$

We will leave out other, practically less relevant combinations such as $(A > B) \vee (B > A)$. This leaves us with six substates of the preference state with respect to the two relata A and B, namely $A > B$, $B > A$, $A \equiv B$, $A \geq B$, $B \geq A$, and $A <> B$. A change of preference concerning the two alternatives A and B consists in a move from one to another of these six states. Such changes group into three different kinds. (We leave aside the trivial type of "change" in which the prior and the posterior states coincide.)

The first type of change concerning A and B occurs when the part of the preference set that refers only to A and B is replaced by one of its proper supersets. This happens when an undetermined state is changed either into a weak preference, a strict preference or a value-equality, or when a weak preference is changed either into a strict preference or a value-equality. We call this type of change an *expansion* with respect to A and B. Figure 8.1 captures the intuitive notion of expansion.

The second type of change occurs when the preference set is replaced by one of its proper subsets. This happens when a value-equality or a strict preference is changed into a weak preference, or when a value-equality, strict preference or weak preference is transformed into an undetermined state. We call this type of change a *removal* with respect to A and B. Figure 8.2 clarifies the inverse relation between expansion and removal.

The third type of change occurs when the preference set is replaced by another, such that neither of them is a proper subset of the other. This happens when a strict or weak preference is changed into a strict or weak preference in the opposite direction,

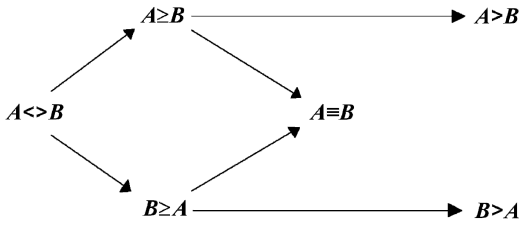

Fig. 8.1 Expansion of a preference state with respect to A and B

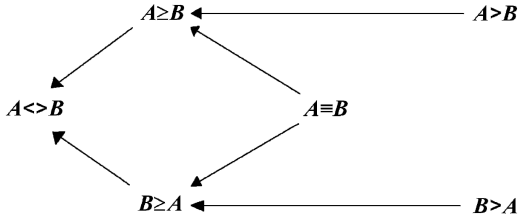

Fig. 8.2 Removal on a preference state with respect to A and B

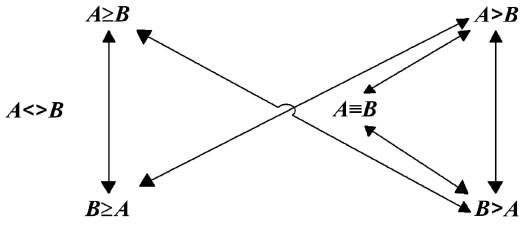

Fig. 8.3 Exchange in a preference state with respect to A and B

or when a value-equality is replaced by a strict preference, or vice versa. We call such a transformation an *exchange* with respect to A and B, since it consists in removing some relations from the preference set and adding others to it. See Fig. 8.3.

Of course, preference changes normally refer to more than just a single pair of relata. More generally, we will call a preference change

a *removal* if it is a removal with respect to all pairs that are affected by the change,
an *expansion* if it is an expansion with respect to all pairs that are affected by the change, and
an *exchange* if it is neither a removal nor an expansion.

This is an exhaustive categorization, i.e. all non-vacuous changes belong to exactly one of these categories. It is important to note that nothing has yet been said here about the input that gives rise to the change.

There is an obvious way to "sentencify" each of these three types of changes (Hanssson 1995):

A *removal* can be constructed as a contraction by some sentence(s).
An *expansion* can be constructed as an expansion by some sentence(s).
An *exchange* can be constructed as a revision by some sentence(s), or as a replacement by some pair of sentences.

However, even if such constructions are feasible it remains to determine whether they are plausible.

8.5.4 The Problems of Sentential Input Representation

As already mentioned, in standard accounts of belief change all inputs are defined in terms of sentences. This feature of belief change theory is far from unproblematic. Actual epistemic agents are moved to change their beliefs largely by non-linguistic inputs, such as sensory impressions. Sentential models of belief change (tacitly) assume that such primary inputs can, in terms of their effects on belief states, be adequately represented by sentences. Thus, when a person sees a hen on the roof (a sensory input), she adjusts her belief state *as if* she modified it to include the sentence "there is a hen on the roof" (a linguistic input).

There are many cases when the causal processes underlying belief formation take the form of accepting sentential information in the way that standard belief revision theory presents it. (Education relies to a large part on that mechanism.) There are also preference changes that can be modelled as caused by accepting sentential information. For example, some people try hard to prefer what they believe has value. If they succeed in their endeavours, the sentences that identify the preferences they aspired to will have contributed to their preference change. Something similar will be the case with preference changes induced by accepting an ideological doctrine (Zaller 1992). Further, when agents adopt preferences from their information about what people of high social standing consume (Leibenstein 1950), the sentential representation of these 'celebrity' preferences will play a causal role. And last, preference sentences will be causally efficacious in those cases where parental example and teaching form children's preferences (Cavalli-Sforza and Feldman 1981).

However, in many important classes of preference change, the actual causes of preference formation are decidedly non-sentential. This is most obvious with respect to *visceral preferences*. An increase in the concentration of the hormone leptin in the blood stream, for example, leads to a reduced desire to eat (Zhang et al. 1994). Other hormonal variations affect human sexual libido (Bullivant et al. 2004). Many new preferences are formed (and older ones lost) with increasing age. Thus was the experience of Shakespeare's Benedick: "but doth not the appetite alter? A man loves the meat in his youth that he cannot endure in his age" [*Much ado about nothing* act II, scene III]. Preferences for a romantic partner can slowly and unnoticeably erode, until the sudden realisation that one 'fell out of love' – as Bertrand Russell describes at his own example (1967, p. 195). Preferences can even be lost or formed through physical damage done to the brain tissue. For example, Oliver Sacks reports a case where an outbreak of neurosyphilis awakened in a shy elderly lady a preference for telling jokes and flirting. In another case, a carcinoma in the brain apparently transformed a reserved research chemist into an impulsive and facetious punster (Sacks 1987, pp. 97–111). In all these cases it is quite obvious that preference sentences are not part of the cause of the described changes. Philosophers of

all ages have acknowledged the importance of these kinds of preference change. "A man often believes himself leader when he is led; as his mind endeavours to reach one goal, his heart insensibly drags him towards another" (La Rochefoucauld 1871, maxim 43). "The heart has its reasons, which reason does not know" (Pascal 1958, Section 277). It seems at least possible that affects sometimes have a direct effect on action (compare the German 'im Affekt handeln'); momentary emotions dominate a person so strongly that they appear to be the only causes of her action.

Yet we want to resist the conclusion that there are two wholly distinct kinds of preference change. Instead, we argue for a unified account by pointing out that in most cases, preference affects are curbed by existing preference states. The ascete, no doubt, feels pangs of hunger, but resists eating and even the motivation to search for food. Some people in monogamous relationships feel sexually attracted to others, but resist the impetus to develop a preference for sex with others. Affects may be a source of motivation, but more often than not, they are filtered through the accepted preferences people hold.

Thus, affects can determine an action directly, or alternatively they can influence the accepted preference state and thereby the individual's actions. Figure 8.4 presents these two effects of preference affects.

This provides us with a cue to how a model of preference change can accommodate the wide variety of decidedly non-sentential causes of preference change, and yet allow for logical analysis. A distinction should be made between the formation of affects and the effects of affects on preference states. Affects can be taken as *primary inputs* that give rise to a *secondary input* in the form of a new preference pattern that has to be incorporated into the preference state.[9] This secondary input has to be expressible in the language of preferences. This gives rise to the structure of preference change theory presented in Fig. 8.5.

It is therefore possible to represent the input by a sentence or set of sentences, although this is of course an idealization. Models of belief change similarly idealize when modelling inputs as sentences, yet the difference between primary and secondary inputs appears to be more important in preference change. Thus, in this way, models of preference and belief change differ.

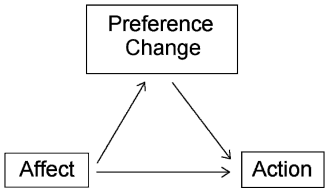

Fig. 8.4 Direct and preference-mediated effects of affects on actions

[9] We will continue to use the term "input" for "secondary input" when that can be done without causing confusion.

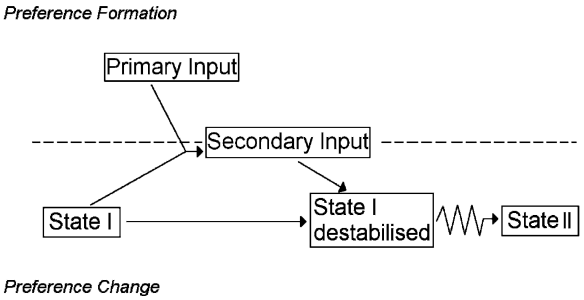

Fig. 8.5 The roles of primary and secondary inputs in preference change

8.5.5 Summary

Preference change, we argue, is exhaustively characterized by three kinds of relationships between prior and posterior preference states: removal, expansion and exchange. For many reasons, however, it is desirable to provide a more structured representation. We therefore propose an input-assimilating model of preference change, in which each change is seen as the reaction to a sentential input. Such a model requires the identification of a sentential input for each kind of preference change. Distinguishing between the formation of preference affects and the subsequent preference change proper allows such an idealized representation.

8.6 Priority-Setting

8.6.1 The Need for Prioritizing Mechanisms

In belief revision, the combination of the integrity constraints and the input constraints is not sufficient to determine the output. As an example of that, consider standard AGM belief contraction. Let the language be infinite. Let the belief set $\mathbf{K}$ be the closure of a single contingent sentence, $\mathbf{K} = \mathrm{Cn}(\{p\})$. Let q be any contingent sentence that follows logically from p. Our task is to contract $\mathbf{K}$ by q. Then there is an infinite number of belief sets that satisfy the integrity and input constraints for this operation, i.e. there is an infinite number of belief sets that satisfy the AGM postulates for being the result of contracting $\mathbf{K}$ by q (Hansson 2008). In view of this, it should be no surprise that belief revision theory makes abundant use of various formal methods to select among the contraction outcomes that are compatible with the integrity and input constraints. The most well-known such formal mechanisms are selection functions and entrenchment relations. In addition, belief bases can be seen as a means to set priorities. If we replace the belief set $\mathbf{K} = \mathrm{Cn}(\{p\})$ by any finite belief base B such that $\mathrm{Cn}(B) = \mathbf{K}$, then of course there are only a finite number

of contraction outcomes that satisfy the integrity and input constraints, given that being a subset of B is one of the input constraints.

Turning to preference change, the need for priority-setting mechanisms is more difficult to assess since we do not have a well-investigated canonical account of the integrity and input constraints as we have for belief change. However, it seems to be a realistic assumption that priority-setting mechanisms are needed here as well.

8.6.2 Priority Information as a (Second-Order) Preference Ordering

In any input-assimilating model, the posterior state is determined by the prior state and the input. Therefore, any priority information will have to be carried either by the prior state or by the input. In standard belief revision theory, all priority information comes with the prior belief state. An entrenchment ordering, for instance, is a prior ordering of the belief set that is one and the same for all inputs. In belief revision, a natural interpretation of the entrenchment ordering is that agents should give up beliefs that have as little explanatory power and overall informational value as possible. As an example of this, in the choice between giving up beliefs in natural laws and beliefs in single factual statements, beliefs in the natural laws, having much higher explanatory power, should in general be retained (Gärdenfors 1988).

For a theory of preference change, the technical correlate to such a priority ordering is an ordering of preferences. The notion of 'second-order preferences' (Sen 1977) or 'preferences amongst preferences' (Jeffrey 1974) has been used to investigate questions of morality, personhood, and akrasia. It can also be reinterpreted for the present purpose. Jeffrey offers the example of the 'good soldier', who prefers adopting his preferences to his orders, rather than following his appetites, fears, or moral judgments, and who thus has a second-order preference ranking for adopting certain sorts of first-order preferences on command (Jeffrey 1974, pp. 158–159). The good soldier thus has an ordering of his preference set that is the same for all inputs and identifies which preferences are to be excised first when in conflict with other preferences.

The usefulness of this notion, however, may be limited. While explanatory power and overall informational value provide an intersubjective criterion by which to interpret the priority ordering for beliefs, the values at the basis of second-order preferences are highly subjective and may shift at any moment. Jeffrey's 'good soldier' is more characteristic of a role that a person can take than of the person himself: at a moment's notice, the 'good soldier' may change into a 'conscientious citizen' or a 'self-preserving egoist'. Considering this instability raises uncomfortable questions about the applicability of second-order preferences to the regulation of (first-order) preference change.

8.6.3 Priority Information in the Preference Base

As we argued in Section 8.3.3, if the preference state is constructed with a preference base instead of a preference set, then there is another way to convey priority information, namely as encoded in the choice of which preferences to include in the preference base. For illustration, imagine two gourmets with identical preference sets, Hans and Peter. Hans prefers Japanese cuisine to Italian, and Italian to French. Consistency also commits him to prefer Japanese to French – even though he never compared the two cuisines. Peter also prefers Japanese cuisine to Italian, and Italian to French. But he, in contrast to Hans, has compared Japanese to French, and found that he preferred the former to the latter. Their respective preference bases look as follows:

Hans:
{*Japanese*>*Italian*, *Italian*>*French*}
Peter:
{*Japanese*>*Italian*, *Italian*>*French*, *Japanese*>*French*}

Hence their preference bases are statically equivalent in the sense that the logical closures of their respective bases are identical. Then, after coming together for a dinner of Italian and French food, the two conclude that in fact the latter cuisine is better than the former, and they both accept a preference for French over Italian. Registering these changes leads to the following preference bases:

Hans:
{*Japanese*>*Italian*, *French*>*Italian*}
Peter:
{*Japanese*>*Italian*, *French*>*Italian*, *Japanese*>*French*}

These bases are no longer statically equivalent: Hans no longer prefers *Japanese* to *French*, which he was previously committed to for reasons of consistency. As soon as his preference base changed, that preference was no longer needed and dropped out of his preference set. Not so for Peter, who had accepted that preference in its own right.

When, after a second dinner, our friends conclude that Italian cuisine is in fact better than Japanese, Peter faces a difficult choice adjusting his preferences, while Hans has no such problem.

Hans:
{*Italian*>*Japanese, French*>*Italian*}
Peter:
{*Italian*>*Japanese*, *French*>*Italian*}
or
{*Italian*>*Japanese*, *Japanese*>*French*}
or
{*Italian*>*Japanese*}

Hans simply changes his preference over *Japanese* and *Italian* (accepting the derived preference for *French* over *Japanese* as a matter of consistency). Given his preference base, that is his unique reasonable choice. Not so for Peter. In order to accommodate his new preference for *Italian* over *Japanese*, he has to give up one or both of his two other preferences.

This example shows how information about the origin of identical preference sets can lead to differences in how these preference sets are adjusted in the face of new preferences. Thus, preference bases contain relevant priority information.

8.6.4 Input-Carried Priority Information

In preference change there are strong reasons to consider whether priority information can be carried by the inputs. The reason for this is that the context of the (secondary) input that gives rise to it (namely the primary input) often contains priority information of a special kind. There are, for instance, two major ways to change one's preferences in order to accommodate a new preference representable as $A \geq B$: either you change the position of A in the preference ordering or that of B. The primary input often tells us which of these to choose: You get tired of brand A, and start to like it less than brand B that was your previous second choice. You learn that the political party X has changed its policies on unemployment insurance, and start to like it more than party Y, etc. (Hansson 2001a, pp. 46–47). This positional information is special to changes pertaining to the position of individual items in an ordering. It is not treated in the standard belief revision models, which focus on changes of membership of individual items in a set. Thus, the need to deal with this special kind of priority information is another feature that sets models of preference change apart from models of belief change.

To exemplify this, consider Mr. Myer who orders four newspapers transitively as follows:

$$A > B > C > D \tag{8.15}$$

Case (i): He learns that newspaper A has participated in a cover-up of severe crimes committed by its principal owner. As a consequence of this, he changes his opinion about A, and now considers it to be worse than D.

Case (ii): He finds out that newspaper D has improved dramatically since he last read it, and now finds it to be even better than A.

Intuitively, in case (i), the outcome of his preference change should be $B > C > D > A$, whereas in (ii) it should be $D > A > B > C$. In case (i), it is A's position that should be moved relative to D and to all other options ranked relative to D. In case (ii), it is D's position that should be moved relative to A and to all other options ranked relative to A.

However, if the (secondary) input is specified only as revision by a preferred sentence, then this difference will be lost since that sentence will be $D > A$ in both cases. This problem is solved by using composite inputs that specify both the input

preference sentence *and* that relatum whose position is to be changed to accommodate the input sentence. Hence a revision by the preference $A \geq B$ can have a dyadic input that specifies not only $A \geq B$ but also for instance that it is the sentence A that is going to be moved around in the ordering, rather than B.

8.6.5 Summary

As in belief revision, theories of preference change require further criteria that help selecting among removal and exchange outcomes compatible with integrity and input constraints. We consider three sources of such priority information: a second order preference ordering, a preference base, and the input. We caution expectations about information from second-order preferences, as the possible approaches tend to yield frameworks that are too unstable. Instead, we suggest the use of exogeneous information from inputs, and in particular the use of endogenous information from preference bases.

8.7 Conclusion

A reader who hoped for a simple translation of belief change methodology to preference change may be somewhat disappointed at this point. We have argued that the general input-assimilating framework from belief change can be transferred, but we have also indicated several modifications that seem to be necessary. The input model has to be complicated with the introduction of a distinction between primary (non-linguistic) and secondary (linguistic) inputs. The method of sentential representation has to be used with somewhat more caution for preferences than for beliefs. Not least, the priority-setting mechanism has to be adjusted, and it seems useful to include some priority-related information in the inputs.

In summary, preference change cannot be successfully pursued as a straightforward application of belief change. It can make use of many concepts and methods from belief change but it is, definitely, a research area with its own specific problems and potentials in need of investigation.

References

Adams, Robert M. 1974. Theories of actuality. *Noûs* 8: 211–231.

Arrow, Kenneth. 1977. Extended sympathy and the possibility of social choice. *American Economic Review* 67: 219–225.

Brogan, Albert P. 1919. The fundamental value universal. *Journal of Philosophy, Psychology, and Scientific Methods* 16: 96–104.

Bullivant, Susan B., Suma Jacob, Martha K. McClintock, Julie A. Mennella, Sarah A. Sellergren, Natasha A. Spencer and Kathleen Stern. 2004. Women's sexual experience during the menstrual cycle: identification of the sexual phase by noninvasive measurement of luteinizing hormone. *Journal of Sex Research* 41 (1): 82–93.

Carlson, Erik. 1997. A note on Moore's organic unities. *Journal of Value Inquiry* 31: 55–59.

Cavalli-Sforza, Luigi Luca and Marcus W. Feldman. 1981. *Cultural transmission and evolution*. Princeton, NJ: Princeton University Press.

Chisholm, Roderick M. 1963. Supererogation and offence: a conceptual scheme for ethics. *Ratio* 5: 1–14.

Chisholm, Roderick M. and Ernest Sosa. 1966. On the logic of "intrinsically better". *American Philosophical Quarterly* 3: 244–249.

Danielsson, Sven. 1997. Harman's equation and the additivity of intrinsic value. *Uppsala Philosophical Studies* 46: 23–34.

Davidson, Donald. 1980. Hempel on Explaining Action. *Essays on Actions and Events*. Oxford: Oxford University Press.

Fermé, Eduardo and Sven Ove Hansson. 1999. Selective Revision. *Studia Logica* 63: 331–342.

Fermé, Eduardo and Sven Ove Hansson. 2001. Shielded contraction. In *Frontiers of belief revision*, eds. Hans Rott and Mary-Anne Williams, 85–107. Dordrecht, The Netherlands: Kluwer.

Fuhrmann, André and Sven Ove Hansson. 1994. A survey of multiple contractions. *Journal of Logic, Language, and Information* 3: 39–76.

Gabbay, Dov M. 1999. Compromise, update and revision. A position paper. In *Dynamic worlds*, eds. B. Fronhoffer and R. Pareschi, 111–148. Dordrecht, The Netherlands: Kluwer.

Gärdenfors, Peter. 1988. *Knowledge in flux. Modeling the dynamics of epistemic states*. Cambridge, MA: MIT Press.

Halldén, Sören. 1957. *On the logic of 'better'*. Lund, Sweden: Library of Theoria.

Hansson, Sven O. 1994. Taking belief bases seriously. In *Logic and philosophy of science in Uppsala*, eds. Dag Prawitz and Dag Westerståhl, 13–28. Dordrecht, The Netherlands: Kluwer.

Hansson, Sven O. 1995. Changes in preference. *Theory and Decision* 38: 1–28.

Hansson, Sven O. 1997. Semi-revision. *Journal of Applied Non-Classical Logic* 7:151–175.

Hansson, Sven O. 1999. *A textbook of belief dynamics, theory change and database updating*. Dordrecht, The Netherlands: Kluwer.

Hansson, Sven O. 2001a. *The structure of values and norms*. Cambridge: Cambridge University Press.

Hansson, Sven O. 2001b. Preference logic. In *Handbook of philosophical logic*, ed. Dov Gabbay and Franz Guenthner, 2nd ed, vol 4, 319–394. Dordrecht, The Netherlands: Reidel.

Hansson, Sven O. 2006. Ideal worlds – wishful thinking in deontic logic. *Studia Logica* 82: 329–336.

Hansson, Sven O. 2008. Specified meet contraction. *Erkenntnis* 69: 31–54.

Hansson, Sven O. 2009. Replacement – a Sheffer stroke for belief revision. *Journal of Philosophical Logic*, 38:127–149.

Harman, Gilbert H. 1967. Toward a theory of intrinsic value. *The Journal of Philosophy* 64: 792–805.

Jeffrey, Richard. 1974. Preferences among preferences. *Journal of Philosophy* 71(13): 377–391. Reprinted in *Probability and the art of judgment*, 154–169. Cambridge: Cambridge University Press, 1992.

Lee, Richard. 1984. Preference and transitivity. *Analysis* 44: 120–134.

Leibenstein, Harvey. 1950. Bandwagon, snob and veblen effects in the theory of consumer's demand. *Quarterly Journal of Economics* 64: 183–207.

Levi, Isaac. 1974. On indeterminate probabilities. *Journal of Philosophy* 71: 391–418.

Levi, Isaac. 1977. Subjunctives, dispositions and chances. *Synthese* 34: 423–455.

Levi, Isaac. 1991. *The fixation of belief and its undoing: changing beliefs through inquiry*. Cambridge: Cambridge University Press.

Makinson, David. 1997. Screened revision. *Theoria* 63: 14–23.

Moore, George. E. 1903. *Principia ethica*. Cambridge: Cambridge University Press.

Oldfield, Edward. 1977. An approach to a theory of intrinsic value. *Philosophical Studies* 32: 233–249.

Olsson, Eric J. 1997. A coherence interpretation of semi-revision. *Theoria* 63: 105–134.

Packard, Dennis J. 1987. Difference logic for preferences. *Theory and Decision* 22: 71–76.

Pascal, Blaise. 1958. *Pensées*. New York: Dutton.

Quinn, Warren S. 1974. Theories of intrinsic value. *American Philosophical Quarterly* 11: 123–132.

Reynolds, James and David Paris. 1979. The concept of choice and arrow's theorem. *Ethics* 89: 354–371.

Rochefoucauld Duc De La, Francois. 1871. *Reflections; or sentences and moral maxims*. Ed. J. W. Willis Bund and J. Hain Friswell. London: Simpson Low, Son, & Marston.

Russell, Bertrand. 1967–1969. *The autobiography of Bertrand Russell*, 3 vols. London: Allen & Unwin.

Sacks, Oliver. 1987. *The man who mistook his wife for a hat*. New York: Harper & Row, Perennial Library Edition.

Sen, Amartya. 1973. Behaviour and the concept of preference. *Economica* 40: 241–259.

Sen, Amartya. 1977. Rational fools. In *Choice welfare and measurement*, 84–106. Cambridge, MA: Harvard University Press, 1982.

Slote, Michael. 1984. Satisficing consequentialism. *Proceedings of the Aristotelian Society*, Supplementary volume 58: 139–164.

Spohn, Wolfgang. 1978. *Grundlagen der Entscheidungstheorie*, Monographien Wissenschaftstheorie und Grundlagenforschung, No. 8. Kronberg/Ts: Scriptor.

Stigler, George J. and Gary S. Becker. 1977. De gustibus non est disputandum. *American Economic Review* 67: 76–90.

Trapp, Rainer W. 1985. Utility theory and preference logic. *Erkenntnis* 22: 301–339.

Ullmann-Margalit, Edna and Sidney Morgenbesser. 1977. Picking and choosing. *Social Research* 44: 757–785.

von Wright, Georg H. 1963. *The logic of preference*. Edinburgh: Edinburgh University Press.

Zaller, John R. 1992. *The nature and origins of mass opinion*. New York: Cambridge University Press

Zhang, Yiying, Ricardo Proenca, Margherita Maffei, Marisa Barone, Lori Leopold and Jeffrey M. Friedman. 1994. Positional cloning of the mouse obese gene and its human homologue. *Nature* 372: 425–432.

Chapter 9
Preference Utilitarianism by Way of Preference Change?[1]

Wlodek Rabinowicz

Abstract This paper revisits Richard Hare's classical and much discussed argument for preference utilitarianism (*Moral Thinking*, 1981), which relies on the conception of moral deliberation as a process of thought experimentation, with concomitant preference change. The paper focuses on an apparent gap in Hare's reasoning, the so-called No-Conflict Problem. A solution to this difficulty which was proposed in (Rabinowicz and Strömberg 1996) is re-examined and shown to lead to a number of difficulties. The paper therefore also considers an alternative idea, due to Daniel Elstein. This new proposal may well turn out to be the best way of filling the gap in Hare's argument.

The paper also examines whether the gap is there to begin with: The problem should perhaps be *dis*solved rather than solved. This suggestion goes back to an idea of Zeno Vendler (1988). Unfortunately, it turns out that Vendler's move does not save Hare from criticism: It does dissolve the No-Conflict Problem, but at the same time it gives rise to another, potentially more serious difficulty.

In this paper, I revisit Richard Hare's classical and much discussed argument for preference utilitarianism (Hare 1981). The argument, which relies on the conception of moral deliberation as a process of thought experimentation, with concomitant preference change, is problematic in several respects. Here, I shall mainly focus on one of these difficulties: on an apparent gap in Hare's reasoning, which might

W. Rabinowicz
Department of Philosophy, Lund University
e-mail: Wlodek.Rabinowicz@fil.lu.se

[1] I am indebted to Daniel Elstein, Christian List, Toni Rönnow-Rasmussen, Mark Schroeder and Bertil Strömberg for very useful comments. Earlier versions of this paper were presented at a workshop on preference change, arranged in connection with the congress of the Gesellschaft für Analytische Philosophie in Berlin 2006, and then at a meeting of the British Society for Ethical Theory, in Edinburgh 2008. I'd like to thank the participants of these events and the organizers: Till Grüne-Yanoff with Sven Ove Hansson and Ellinor Mason with Michael Ridge, respectively. Last, but not least, I wish to thank the referees for this volume, Peter Dietsch and Martin Peterson, who both have kindly agreed to disclose their identity. Their suggestions have been very helpful.

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

be called The No-Conflict Problem. In a paper I wrote with Bertil Strömberg several years ago, we tried to fill this lacuna (Rabinowicz and Strömberg 1996). Our suggestion focused on the idea that moral deliberation requires preference revision: One's preferential state needs to be revised so as to satisfy a certain uniformity constraint. Preference revision has therefore to be given a systematic account. We suggested that such revision can be assumed to be guided by the principle of minimal change: The output state must satisfy the imposed constraint (in our case, the constraint of uniformity), but it should otherwise differ as little as possible from the input state. If the measure of distance between preference states is then chosen in the right way, the output state can be shown to be in conformity with utilitarianism. This proposal, however, turns out to lead to a number of difficulties: The choice of an appropriate measure of distance between preference states itself poses a serious problem and it is not even clear that the principle of minimal change, which has often been assumed as the leading principle of *belief* revision, can be justified when it comes to revision of preferences. These, and other difficulties, put our original proposal into doubt. I shall therefore also consider an alternative solution, which was recently suggested to me by Daniel Elstein. The latter proposal is closer to Hare's own way of reasoning and it may well turn out to be the best way of filling the gap in his argument.

In my paper with Strömberg, we also examine whether the gap is there to begin with: The problem should perhaps be *dis*solved rather than solved. This suggestion goes back to an idea of Zeno Vendler (1988). Unfortunately, it turns out that Vendler's move does not save Hare from criticism. It does dissolve the No-Conflict Problem but at the same time it gives rise to another, potentially more serious difficulty.

9.1 The Argument and the Gap

Hare's argument rests on his interpretation of moral judgments as *universal overriding prescriptions*.[2] By the principle of universalizability, a moral prescription that concerns a given situation also applies to the hypothetical variants of that situation, in which the individuals' roles have been reversed. To reach a moral judgment regarding the situation at hand, I must therefore take all these variants into consideration. Thus, suppose I contemplate an action that, apart from me, concerns some other persons, say, John and Mary. The judgment that I ought to perform the action would amount to prescribing it both for the situation I am in *and* for the hypothetical situations in which I would be at the receiving end instead. Consequently, to reach a moral judgment, I have to ask myself: What if I were in John's or Mary's shoes? How would it be like to be subjected to this action? Because of their universal application, moral judgments must be based on, or tested by, this kind of thought-experiments.

[2] See, for example, Hare (1981), p. 55.

"Being in somebody else's shoes" is a somewhat misleading expression in this context. Universalizability only commits me to extending my prescriptions from a given situation to its exactly similar variants. As Hare puts it:

> [I]f I now say I ought to do a certain thing to a certain person, I am committed to the view that the very same thing ought to happen to me, were I in exactly his situation, including having the same personal characteristics and in particular the same motivational states. But the motivational states he actually now has may run quite counter to my own present ones. (Hare 1981, p. 108)

Therefore, when I imagine being as John is now, I must assume that, in this hypothetical situation, I not only take over John's external circumstances but also his body, his psychological make-up, his character, beliefs, emotions and desires – not just the shoes but also what they are sitting on. I try to imagine how it would be like to be subjected to the action if I were just as he is in the actual situation.

To make my discussion less abstract, let me add some detail to the example. Suppose I have agreed to meet John and Mary, two of my students, at the department today. We haven't fixed any definite time for the meeting but the secretary phones me at home with the message that the students have already arrived and are waiting. Since the weather is nice, I would much prefer to go by bike to the office rather than to drive. The students, on the other hand, dislike waiting: They would prefer that I arrive as soon as possible. The preference-utilitarian solution would prescribe the action that best satisfies the balance of our preferences: My preference for going by bike is weighed in proportion to its strength against the students' opposing preferences. Suppose that each of the latter is weaker than my own but that together they weigh more. Under these circumstances, I ought to abstain from going by bike and take the car instead.

Balancing pre-supposes that the strength of people's preferences can be compared and measured on a common scale. Obviously, this is a highly contentious claim, but, as Hare takes it more or less for granted (see Chapter 7 in Hare 1981, esp. 124), I am going to do likewise, at least for the argument's sake. Suppose then that the strength of my preference for going by bike is $+4$, while the intensities of John and Mary's opposing preferences are -3 and -2, respectively. The signs, plus or minus, specify the direction of a preference – whether it is for or against the action under consideration. The preference-utilitarian calculus implies that I should abstain from the bike alternative: $+4 - 3 - 2 < 0$.

Now, Hare wants to establish that I will reach the same result if I seek to arrive at a moral judgment *via* thought-experiments in which I take on the positions of the different persons involved in the situation at hand. If I proceed in this way, and if I am rational, well-informed and equipped with sufficient imagination, I cannot avoid arriving at the same moral prescription as the one delivered by preference utilitarianism.

How does Hare describe the process that leads me towards a moral judgment? Let me start with a quote from *Moral Thinking*. There, he discusses a 'bilateral' example in which I – the subject – consider whether to move someone else's bicycle in order to create a parking space for my car. No other persons are involved. Preference utilitarianism implies that I ought to move the bicycle if, and only if, my preference

for this action is stronger than the cyclist's preference against his bicycle being moved. Hare comments:

> I can see no reason for not adopting the same solution here as when we do in cases when our own preferences conflict with one another. For example, let us change the case and suppose that it is my own bicycle, and that it is moderately inconvenient to move it but highly inconvenient not to be able to park my car; I shall then naturally move the bicycle thinking that that is what, prudentially speaking, I ought to do, or what I most want, all in all, to do. Reverting now to the bilateral case: we have established [Section 5.3, pp. 94–96] that, if I have full knowledge of the other person's preferences, I shall myself acquire preferences equal to his regarding what should be done to me were I in his situation; and these are the preferences which are now conflicting with my original prescription [to move the bicycle]. So we have in effect not an interpersonal conflict of preferences or prescriptions, but an intrapersonal one; both the conflicting preferences are mine. I shall therefore deal with the conflict in exactly the same way as with that between two original preferences of my own.
>
> Multilateral cases [in which several persons are affected] now present less difficulty than at first appeared. For in them too the interpersonal conflicts, however complex and however many persons are involved, will reduce themselves, given full knowledge of the preferences of others, to intrapersonal ones. (Hare 1981, p. 109f)

Let us try to unpack this passage, now using the example that we have started with. I contemplate going by bike to the office in the situation at hand, call it s_1. I have a preference *for* this action, with strength 4. However, since moral judgments are universal, they prescribe exactly similar things for exactly similar situations. Consequently, a moral judgment concerning what I ought to do in s_1 would also apply to the hypothetical situations in which the roles were reversed. Therefore, I need to imagine being in John's shoes and in Mary's shoes, respectively, i.e. to envision two hypothetical situations, s_2 and s_3, in each of which I am on one of the receiving ends. I realize that if I were in John's position, with his desires etc., I would have the same preference as John has in the actual situation: *against* the action in question, with strength 3. Analogously, were I in Mary's shoes, I would have a preference *against* the action, with strength 2.

The next step in the deliberation process pre-supposes what Allan Gibbard has called the Principle of Conditional Reflection (Gibbard 1988). Hare himself introduces that principle without giving it any label.

Conditional Reflection: Insofar as I fully know what I would prefer *in* a hypothetical case, I must have the corresponding preference (same sign, same strength) *with regard to* that hypothetical case.[3]

In other words, my hypothetical preferences – if I know I would have them in a hypothetical case – are reflected in my actual preferences with regard to the case in question. Insofar as I now come to see that I would disprefer the biking alternative if I were in John's position, I acquire a preference against this action with regard to that hypothetical situation.

[3] Cf. Hare (1981), p. 99: "I cannot know the extent and quality of others' suffering and, in general, motivations and preferences without having equal motivations with regard to what should happen to me, were I in their places, with their motivations and preferences." The same principle has also been called "The Principle of Hypothetical Self-Endorsement" (in Persson 1989), and "The Principle of Conditional Self-Endorsement" (in Rabinowicz 1989; Rabinowicz and Strömberg 1996).

Hare takes Conditional Reflection to be a conceptual truth. The principle holds due to the alleged presence of a prescriptive element in the very concept of 'I': "The suggestion is that 'I' is not wholly a descriptive word but in part prescriptive" (ibid., p. 96).

> [B]y calling some person 'I', I express at least a considerably greater concern for the satisfaction of his preferences than for those of people whom I do not so designate. Thus, in a normal clear-cut case, if I were asked, when somebody is being maltreated and dislikes it, 'How do you feel about being put forthwith in that situation with his preferences', I shall reply that if it would be *me*, I do now have the same aversion to having it done as he now has. (ibid., p. 98)

Conditional Reflection is grounded in my fundamental self-concern, which Hare interprets as a concern for the satisfaction of my preferences, whether actual or hypothetical. In thinking of hypothetical preferences as *mine*, I thereby endorse them. Is it a convincing claim? One might doubt this: Such endorsement seems suspended when I consider hypothetical cases in which my preferences by my *present* lights would be corrupted or distorted in some way. If I had a sadistic disposition, I would wish to cause pain. But knowing this doesn't make me wish to cause pain if I were a sadist: I do not endorse my hypothetical preference if I now judge it to be corrupted or irrational.

This suggests that Conditional Reflection should at least be qualified in some ways in order to be acceptable. Also, perhaps it should be interpreted as a requirement of rationality rather than as a conceptual truth: While self-concern that underlies Conditional Reflection seems to be an important element of rationality, one might well question whether this attitude plays a role in determining the very meaning of terms such as "I" or "mine". If viewed as a rationality requirement, Conditional Reflection may be interpreted as a condition on ideally self-integrated and self-confident preferrers.[4] Still, at least for the time being, we can leave this principle as it stands. It is in any case clear that the "distortion" objection does not apply in our example: The students' preferences for my early arrival are perfectly reasonable.

Conditional Reflection implies that, after having considered what it would be like to be in my students' shoes, I end up with several preferences as regards the contemplated action – as many as the number of the situations I have considered.

[4] As is easily seen, Conditional Reflection, which is a constraint on preferences, is closely related to the well-known reflection principle for beliefs. According to the latter, knowing what one would believe in a hypothetical situation commits one to analogous and equally strong conditional beliefs – conditional on the obtaining of the situation in question. Bas van Fraassen has shown that a person who violates that principle is vulnerable to a Dutch Book, provided only that she assigns some positive probability to the hypothetical situation in question (cf. van Fraassen 1984). In Rabinowicz (1989), I suggest that a similar Dutch Book argument might be available for an analogous reflection principle for preferences. However, as Hare would readily admit, the probability of my occupying exactly the same position as someone else occupies in a situation at hand is zero. Therefore, it isn't possible to set up a Dutch Book against someone who violates Conditional Reflection with regard to such hypothetical cases. If Conditional Reflection is to be defended even for these thought-experiments, the defense could not proceed on such purely pragmatic grounds.

I still have my original preference for the bike alternative with strength 4, but now – after having considered the hypothetical situations s_2 and s_3 – I also acquire two preferences against this action, with strengths 3 and 2, respectively.

In the passage quoted above, Hare seems to suggest that the last step in the process of arriving at a moral judgment consists in prudential *balancing*. Here I am, with preferences that pull me in opposing directions – towards the action and away from it. My rational preference "all in all", as he puts it, is a function of these preferential inputs. In our example, this means that I come to prefer not to go by bike, all in all: My preferences against this action are jointly stronger than my preference for the action. By engaging in thought-experiments that lead me to acquire new preferences and then by balancing these against my original preference, I thus seem to have reached the same solution as the one delivered by preference utilitarianism. Hare is anxious to point out that his re-construction of the process of moral deliberation transforms the original *inter*personal preference conflict into a conflict that is *intra*personal. The latter is then solvable in the standard way – by simple balancing.

Now, as Schueler (1984) and Persson (1989) have pointed out, this argument – as presented above – contains an important gap. Hare's comparison with standard decision problems in which the subject experiences a conflict of preferences is misleading. In the standard case, my conflicting desires concern one and the same situation – the one in which I make my choice. In Hare's argument, however, the various preferences I have acquired via thought-experiments are not related in this way. I have a preference for going by bike with regard to the actual situation s_1, a preference against this action with regard to the hypothetical situation s_2, in which I am in John's shoes, and yet another preference against this action with regard to s_3, in which I occupy Mary's position. These desires of mine concern different situations and for that reason they do *not* oppose each other. Unlike as in the prudential case, there is here no conflict of preferences to begin with, which would need to be solved by balancing. Thus, suppose I would decide to go by bike to the office. This action would satisfy my preference as regards the actual situation s_1, but it would in no way frustrate my preferences regarding the purely hypothetical situations s_2 and s_3. This, in a nutshell, is the 'No-Conflict Problem' that threatens Hare's argument.

But haven't we forgotten something? What about the principle of universalizability? Universalizability requires that my prescription with regard to the different situations under consideration, s_1, s_2 and s_3, must be *uniform* in order to be moral. Thus, as long as my preferences regarding s_1 differ from those regarding s_2 and s_3, I haven't yet arrived at a moral judgment. While it is not made clear in the above quoted passage, Hare elsewhere seems to suggest that the uniform prescription can be reached by a process of *tentative extrapolation*: I try to extrapolate my preference regarding one situation to other situations. The question is then whether the extrapolated preference is strong enough to survive any conflicts of preference that might be created by this move.[5] If it is not, then I try to extrapolate one of my other

[5] "if I now say that I ought to do a certain thing to a certain person, I am committed to the view that the very same thing ought to be done to me, were I exactly in the same situation. But [. . .] he may very much want not to have done to him what I am saying I ought to do to him [. . .] [I]f

preferences instead – one of those I entertain with regard to the situations in which the roles are reversed. Can this tack help us here?

The extrapolation manoeuvre does help, but only in *bilateral* cases. If there is just one student, say John, who is waiting for me at the department, I just have two situations to worry about, the actual situation s_1 and the hypothetical situation s_2, in which I am in John's shoes. Then I can successfully extrapolate my preference for going by bike from s_1 to s_2, since this preference is stronger than my opposing preference regarding s_2, which I have acquired in accordance with Conditional Reflection. Had the latter preference been stronger, then I would have been able to successfully extrapolate that preference instead. Consequently, I can uphold a uniform prescription with regard to both situations and the prescription is going to be of the right utilitarian kind.

However, the proposed solution would lead us astray in multilateral cases (cf. Persson 1989). Thus, consider again the example with two students who wait for me in the department. It is easily seen that my preference for biking as regards s_1 can be successfully extrapolated to both s_2 and s_3. It is stronger than each of the opposing preferences I have regarding these two situations, even though it is weaker than both of them taken together. The extrapolated preference wins because it only meets one opposing preference at a time. Opposing preferences never have the opportunity to join forces, so to speak. The uniform prescription to go by bike therefore remains undefeated, despite its counter-utilitarian character.

Persson (1989) suggests that the gap in Hare's argument might instead be filled in by introducing a "veil of ignorance" – a device that has been made famous by John Rawls and John Harsanyi. Persson's veil of ignorance is the same as Harsanyi's in the latter's "equiprobability model" (see Harsanyi 1953, 1977): After having acquired preferences concerning the three situations s_1, s_2, and s_3, I should now pretend that I am uncertain as to which of these three situations is the actual one. As Harsanyi suggests, uncertainty should be represented as assignment of equal probabilities (rather than as full ignorance, i.e. absence of probabilities, as Rawls would have it). Thus, I should treat the three situations as though they were equiprobable. Persson's next step, just as Harsanyi's, is to apply the standard principle of expected utility maximization in order to identify the action to be performed.[6] In our example, this means that I should abstain from going by bike to the office. This action

I fully represent to myself his situation, [...] I shall myself acquire a corresponding motivation, which would be expressed in the prescription that the same thing *not be* done to me, were I to be forthwith in just that situation. But this prescription is inconsistent with my original 'ought'-statement, if that was, as we have been assuming, prescriptive. [...] I can avoid this 'contradiction in the will' (cf. Kant 1785, p. 58) only by abandoning my original 'ought' statement, given my present knowledge of my proposed victim's situation" (ibid., 108f). Admittedly, this passage is not crystal-clear. While my original preference is first universalized, i.e. extrapolated to the situation in which I am at the receiving end, and while that universal preference is subsequently abandoned to avoid 'the contradiction in the will', Hare does not explicitly state that the attempt at preference universalization is aborted in this case because the opposing preference is stronger.

[6] Harsanyi thought that such pretence of ignorance was appropriate for an ideal observer. It is more difficult to understand how such pretence could even in principle be possible in the context of

would satisfy my preferences if s_1 is actual, but it would frustrate them if one of the other two situations obtains instead; which is twice as probable, given my pretence of ignorance. Just as Harsanyi's model, this proposal delivers the standard utilitarian solution.

Persson's proposal has not been adopted by Hare, despite the fact that, just like Rawls and Harsanyi, Hare also wants to ground morality in the idea of rational choice. However, unlike these two, Hare is anxious to avoid any elements of pretence, of make-believe in his rational reconstruction of moral reasoning.As Persson himself points out, the veil-of-ignorance approach would represent an alien element in Hare's thought, given Hare's project to base ethics on fully rational grounds:

> [T]he addition of PEP [= The Principle of Equal Probability] to Hare's premisses appears highly problematic, since while rationality [. . .] demands that preferences be formed on the basis of all relevant information available to one, PEP requires one to seal oneself off from certain pieces of information (concerning the numerical identity of the particulars involved). (Persson 1989, p. 170)

One might also put it like this: pretence in, pretence out. With premises we only pretend to accept, we would only pretend to accept the conclusion.[7] Therefore, it is no wonder that in his comments on Persson's paper, Hare tries to fill the gap in a different way (cf. Hare 1989). To save space, I won't discuss that proposal. Let me just say I find it quite unsatisfactory.[8]

9.2 Preference Revision

Instead, let me move to the proposal outlined in my paper with Strömberg. Go back to the point at which my thought experiments have led me to acquire a set of preferences concerning the three situations, s_1–s_3. My preference profile at that stage with respect to the action under consideration can be represented by a vector,

$$(+4, -3, -2) \tag{9.1}$$

in which the first component specifies the strength of my preference concerning s_1, the second component specifies the strength of my preference concerning s_2, and so on. The signs, plus or minus, specify the direction of a preference – whether it is for

practical moral deliberation: When I deliberate whether to perform an action, I cannot at the same time pretend that, for all I know, I might be at the receiving end of the action I consider to perform!

[7] It might be objected that even Hare's own approach involves an element of make-believe. After all, don't thought-experiments play a central role in his account of moral deliberation? This, however, would be a misunderstanding. On Hare's account, the subject is asked to consider what would be the case if the roles were reversed. But he is not asked to engage in pretending that, for all he knows, this hypothetical situation might be *actual*. It is here that Hare parts ways with Rawls and Harsanyi.

[8] For an exposition and critical discussion, see Rabinowicz and Strömberg (1996), Section 9.3.

or against the contemplated action. On the basis of this profile, I must now arrive at a moral judgment, i.e. to a universal prescription, either to go by bike or to abstain, which must be the same for all three situations.

The main idea behind our proposal may be formulated as follows: The universal prescription to be reached should agree as much as possible with the subject's original preference profile. This idea can be made more precise in several ways, two of which we outline in our paper. We distinguish between what we call the "preference revision" approach and the "final verdict" approach. Here I will only focus on the former.

Prescribing and preferring are for Hare essentially the same thing. "[A]ll prescriptions, including moral ones, are expressions of preferences or of desires in a wide sense" (Hare 1981, p. 185, cf. also p. 107).[9] Thus, when I try to arrive at a uniform prescription for the three situations in our example, what I am after is a uniform preference with regard to these situations.[10] In other words, I try to revise my original preferences, which differ with respect to the three situations, in order to reach a new preference state with a uniform profile:

$$(x, x, x) \tag{9.2}$$

In this vector, the same (positive or negative) value appears at each place. In the preference state I am after, I have exactly the same preference, for or against the action, concerning each of the three situations, $s_1 - s_3$.

How should I proceed in order to modify my preferences in this way? What is the appropriate value for x?

Preference revision may be seen as a process analogous to revision of *beliefs*. As the ruling principle of belief revision one usually takes the Principle of Minimal Change. When I have to change my beliefs in order to make room for new information, or – more generally – in order to get them in line with some constraint I need to satisfy, I should be conservative. My new beliefs should deviate as little as possible (given the constraint they need to satisfy) from the beliefs I have started with. To put it differently, the distance between my old beliefs and my new beliefs ought to be minimized given the task at hand. Cf. Gärdenfors (1988), p. 8:

> [W]hen evaluating changes of belief, we require that the change be the *minimal* one needed to accommodate the epistemic input that generates the change.

If Minimal Change is also taken as the ruling principle for revision of preferences (which is a big "if"; this assumption will be discussed below), then it follows that the uniform preference state to be reached by the subject should diverge as little as

[9] See also Hare (1987), p. 73: "To want something to happen is to be in a state of mind which, if it had to be put into words, could be expressed by saying that one accepts the prescription that it happen." Ultimately, this idea goes back to Hare (1963), Section 9.4.

[10] "To accept a universal prescription" is consequently the same as "to form a universal preference" (Hare 1989, p. 172). This identification of a universal prescription with a universal preference was also assumed in the extrapolation manoeuvre.

possible from his original preference state. To paraphrase Gärdenfors, we require that the change of preferences be the minimal one needed to satisfy the uniformity constraint that necessitates the change. Thus, the value for x should be chosen in such a way that the distance between the two preferences states be minimized.

But how are we to determine such distances? If one represents preference states as vectors, as we have done, then each state may be seen as a point in a vector space. A space point is describable by its numerical coordinates, which specify its position in each spatial dimension. In our example, we work with three-place vectors, i.e. with points in a three-dimensional space. Generally speaking, the number of dimensions is determined by the number of situations I – the subject – need to consider, i.e., ultimately, by the number of persons involved in the actual situation. For each person, I am to consider a situation – actual or hypothetical – in which I would be in that person's shoes. Had the number of persons involved been smaller, say, just me and John, only two situations would have to be taken into account, instead of three. My preference state would then be representable as a point in a two-dimensional space.

What measure of distance between vectors is it appropriate to accept? It is clear that this measure should be 'impartial'. In particular, it should not favour the subject's preference with regard to the actual situation, s_1, at the expense of his preferences with regard to hypothetical situations, s_2 and s_3. Such partiality would clearly go against the spirit of universalizability that inspires Hare's enterprise. Thus, we take it that universalizability makes its appearance at two places in Hare's argument: first, as a uniformity constraint on the posterior preference state – as a demand that the posterior preference with regard to each situation be the same whatever position one is supposed to occupy in the situation in question; second, as an impartiality constraint on the distance measure – as a requirement that the distance between points in a vector space be invariant under permutations of dimensions.

Consider an n-dimensional space of preference states. As we already know, n is the number of situations to be considered, i.e., at bottom, the number of persons involved.*If* the distance measure on that space is supposed to be of the standard Euclidean type, one that we are used to deal with in other contexts, then the distance between two preference states, $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$, equals the square root of the sum-total of the squared differences between the corresponding components of v and w:

$$\text{Euclidean distance} : [\Sigma_{i=1,\ldots,n}(v_i - -w_i)^2]^{1/2} \tag{9.3}$$

This makes our task solvable: We can determine what value x must take if the Euclidean distance between the prior preference state and the posterior uniform state $(x, \ldots, x)$ is to be minimized.

It can be proved that Euclidean distance is minimized if x is the *average* of the values in the original preference profile.[11] This averaging solution is, of course, very much in the spirit of preference-utilitarianism: The average of the preferences I have acquired, in accordance with Conditional Reflection, with regard to the situations in which I occupy the positions of different individuals equals the average of the preferences these individuals entertain in the actual situation. And preference utilitarianism implies that the action ought to be performed if and only if the latter average is positive.[12]

Thus, in our example, the average of my preferences in the state

$$(+4, -3, -2) \tag{9.4}$$

equals $-1/3$. Consequently, if the right measure for the distance between preference states is Euclidean, the revised uniform state would be:

$$(-1/3, -1/3, -1/3) \tag{9.5}$$

This means that the moral prescription is to abstain from going by bike, just as preference utilitarians would have it.

9.3 Questions

There are, of course, several controversial elements in this proposal. Here are some of the questions that would need to be examined:

(i) Questions about *prescriptions*: Is prescribing really the same thing as preferring? This seems to presuppose a rather simplistic view of our mental life. According to philosophers like Michael Bratman (see Bratman 1987), we should carefully distinguish between such mental phenomena as desires and intentions. One might equally well argue, I guess, that preference and acceptance of a prescription are distinct mental states. This would create problems for the proposal, which follows Hare in his treatment of moral prescriptions as universal preferences.
(ii) Questions about *minimal change*: Is the analogy between belief revision and revision of preferences justified? The principle of minimal change does not seem to be as plausible in the latter case as in the former. In the case of beliefs, conservatism

[11] For the proof, see Rabinowicz and Strömberg (1996).

[12] This holds if the choice problem is binary, i.e., if the only alternatives are performing the action or abstaining. In such a choice, we may assume that positive preference for an action is mirrored by an equally strong negative preference for its alternative. For a discussion of choices between several action alternatives, see the next section. Note also that it doesn't matter whether we go by the average or by the sum-total of preferences, as long as we only consider situations with a fixed number of persons involved. In this paper, we only consider cases in which the set of individuals to be taken into account is given *ex ante*.

in adding beliefs is grounded in the requirement of epistemic responsibility and conservatism in giving up beliefs has its source in the following consideration: *Ex ante*, stopping to believe amounts to an epistemic loss, since one would then stop believing what one currently takes to be *true*. In the case of preference revision, a corresponding justification for the reluctance in relinquishing preferences is not available, unless preferences are interpreted on cognitivist lines, as beliefs that certain objects (= objects of preference) are valuable. But a cognitivist account of preferences is highly questionable. An alternative justification for conservatism might possibly be found, though, in a principle of mental economy. Changing preferences is not easy, and larger changes might be more difficult to bring about than smaller ones.

There is another possible reason for conservatism in preference change, which specifically applies to Hare's account of moral deliberation, but that reason does not sit well with the proposal that currently is under consideration. As we remember, the preferences in the input state are being entertained by the subject according to the principle of Conditional Reflection. But then, if that principle is supposed to be conceptually true, as Hare insists, it becomes something of a mystery how such preferences can ever be changed.[13] Consequently, the move to a new preference state, with uniform preferences, becomes difficult to account for. This difficulty could be avoided if we instead re-interpret Conditional Reflection as a normative constraint rather than as a conceptual one – as a requirement of rationality, which under some circumstances we might tinker with in order to satisfy other constraints, such as universalizability. Another alternative would be to think of preferences in the input state as somehow still residually present even after the move to the output state has been effected. But it is quite unclear what this would mean.

(iii) Questions about choices between *several alternative actions*: Often, the agent's choice problem isn't simply whether to perform a given action or to abstain. Instead, the task is to choose an action from a set of several alternatives. The subject's preferences with respect to different situations are then representable by a *matrix* rather than by a single vector: If the number of situations to be considered is n and the number of action-alternatives is m, a preference state can be represented as a matrix with n columns (one for each situation) and m rows (one for each action). The numerical values in different cells of the matrix specify preference intensities. Thus, the value that appears in the jth row in the ith column specifies the strength of the agent's preference with regard to action j, as concerns situation i. We may suppose that preference strength is measured on an interval scale. Values in the matrix will then be invariant up to positive linear transformations. (We no longer need to determine whether the subject is for or against an action. It is enough to determine whether he prefers or disprefers that action to other alternatives, and by how much. Thus, the zero point of the scale may now be arbitrarily chosen.)

By universalizability, in order to take a moral stand, the subject needs to move from his prior preference state to a new one, representable by a matrix with *uniform rows*.

[13] I am indebted to Mark Schroeder for this objection.

That is, for every action j, the row for j in the new matrix must have the same preference values for each situation: $(x_j, x_j, \ldots, x_j)$. In addition, by the principle of minimal change, the new matrix should deviate from the original one as little as possible. There is a natural way to generalize the Euclidean measure to distances between $n \times m$-matrices: We let the distance between two matrices be the square root of the sum-total of the squared differences between the values in the corresponding cells of the matrices in question. It can be shown that this distance is minimized if and only if the preference value for each action j in the output matrix is the average of the preference values in j's row in the input matrix.[14] This means that the action that ought to be performed is the one that maximizes the average degree of preference satisfaction for the persons who are involved in the situation at hand, just as preference utilitarians would have it.

In what follows, however, I shall for simplicity's sake revert to binary choice problems, in which there is just one action that the agent has to care about. Thus, instead of distances between matrices, we only need to consider distances between vectors.

(iv) Questions about *distance measure*: Why suppose that the correct measure of distance must be Euclidean? Obviously, it's just one possibility among many. What are then the adequacy criteria for a 'reasonable' measure of distance between preference states? We have mentioned one such criterion: impartiality. Another plausible criterion is that the distance between two preference states, v and w, should be an increasing function of the absolute differences between the corresponding preference components in v and w. But these constraints by themselves do not take us very far.

The simplest distance measure one might want to use in this context is the so-called "city block"-distance, which goes by the sum-total of the absolute differences between vectors v and w on each of the n dimensions[15]:

$$\text{City-block distance} : \Sigma_{i=1,\ldots,n} |v_i - w_i| \tag{9.6}$$

Such a measure, however, does not always yield a unique solution for the distance minimizing task. In fact, in the two-dimensional case, the averaging solution is only one of the infinitely many that are possible. If the original vector is of the form (v_1, v_2), then any x between v_1 and v_2 will fill the bill. Thus, for example, if the prior preference state is $(+3, -2)$, the averaging solution would be $(+1/2, +1/2)$. But the uniform vectors that minimize city-block distance from the vector $(+3, -2)$ form a continuum that ranges from $(+3, +3)$ to $(-2, -2)$. If the number of dimensions is larger than two, using city-block distance might sometimes deliver a unique

[14] For the proof, see Rabinowicz and Strömberg (1996).

[15] The name derives from the fact that, from the Euclidean perspective, this measure gives as the distance between two points the length of the shortest path from one point to the other that at each point moves in parallel to one of the axes. It is as though we were constrained to travel between the points along city streets that form a regular cross-pattern.

solution to the minimizing task, *but* there is no guarantee that this solution will be the averaging one. Here is an example in three dimensions: Suppose that the original vector is $(+6, 0, -3)$. The uniform vector that minimizes city-block distance from $(+6, 0, -3)$ is $(0, 0, 0)$, while the averaging solution would be $(+1, +1, +1)$. (Note, by the way, that $(0, 0, 0)$ would still minimize city-block distance if we replaced the first component in $(+6, 0, -3)$ by any value higher than 6.) The conclusion is that if we opt for the city-block as our distance measure, the argument for preference utilitarianism doesn't go through. So, why is the Euclidean measure to be recommended?

These two measures, city-block and the Euclidean distance, are members of a large family of distance measures, all of which have the form:

$$\text{Minkowski distance} : [\Sigma_{i=1,\ldots,n}|v_i - w_i|^k]^{1/k} (k \geq 1) \tag{9.7}$$

If the coefficient k equals 1, we get the city-block; if it is 2, we obtain the Euclidean distance; and so on. The higher k is, the greater weight is given to larger absolute differences between corresponding vector components, as compared to smaller ones. Only when k equals 1, as in the city block, all the differences between the components are weighted equally, independently of size. But already for $k = 2$, as in the Euclidean measure, the larger absolute differences are given a disproportionately greater influence, by exponentiation, as compared with the smaller differences.[16]

Now, to give greater weight to larger differences between the corresponding components of preference states looks very much like a consideration of *fairness*: One thereby disfavours posterior states that in many of their components deviate very little from the corresponding preferences in the prior state, but in some few cases deviate a lot. That is, ultimately, one disfavours posterior states that show small deviations from the preferences of many persons involved in the situation at hand, but for a few persons deviate a lot. In other words, one thereby disfavours sacrificing some to the benefit of many. This gives rise to a puzzle. It is notorious that fairness considerations are alien to the utilitarian outlook. For a preference utilitarian, the only thing that matters is that the overall degree of preference satisfaction is maximized (the average or the sum-total; it doesn't matter which as long as the population is kept fixed). Whether this goal is accomplished by letting the preferences of some individuals be sacrificed to the benefit of others is irrelevant: Achieving a fair distribution of preference satisfaction doesn't matter. So how can we explain that it is the Euclidean measure rather than the city-block that gives us the utilitarian averaging solution, if it is the former and not the latter that makes allowance for the considerations of fairness? I wish I knew the answer to this puzzling question.[17]

(v) Questions about *Harean exegesis*: How faithful is this proposal to Hare's own formulation of his argument? Well, apart from the obvious fact that Hare never

[16] At the limit, all weight is placed on the largest difference. A very simple distance measure that is sensitive only to the largest differences can be defined as follows: the distance between v and $w = \max\{|v_i - w_i| : i = 1, \ldots n\}$.

[17] I am indebted to Christian List for pressing this point.

considers the problem in terms of minimization of distance between preferential states, there is one big difference between the two approaches: They implement the universalizability requirement in different ways. Preference extrapolation, which in Hare's argument functions as a universalization device, never comes into play in the proposal we have now been considering. Instead, it is replaced by preference revision, in which universalizability is implemented in two ways: as the uniformity constraint on the outcome of revision and as the impartiality constraint on the measure of distance. This also means that Hare's idea of arriving at moral prescriptions by transformation of interpersonal preference conflicts into intrapersonal ones is not preserved.

9.4 Simultaneous Extrapolation

Can we reconstruct the argument in a way that is closer to the original? Let's again go back to the point at which I entertain a set of preferences with varying strengths and signs regarding a given action: one with respect to the actual situation and the remaining ones with respect to the hypothetical situations in which the roles are reversed. As we remember, Hare's suggestion was that a uniform prescription can be reached at that point by a process of tentative extrapolation: I try to extrapolate my preference concerning, say, the actual situation to its hypothetical variants. If the extrapolated preference is strong enough to survive any conflicts of preference that might be created by this move, then I am home. If it is not, then I try to extrapolate one of my other preferences instead – one of those I hold with respect to the situations in which the roles are reversed. As we have seen, however, this proposal can only deal with bilateral cases: It leads to unwelcome results when several persons are involved.

An alternative would be to employ what might be called a *simultaneous preference extrapolation*. This suggestion is due to Daniel Elstein.[18] Let's illustrate how this procedure is supposed to work in our example. I have acquired a preference regarding the action under consideration with respect to each of the situations s_1–s_3. These preferences form a vector,

$$(+4, -3, -2) \tag{9.8}$$

To satisfy the universalizability requirement, I now simultaneously extrapolate each of the preferences in this profile to all the three situations. We can think of this step as a move in which each preference I have is universalized, so as to become a moral prescription. I thus arrive to a complex preferential state in which each of the preferences in the state $(+4, -3, -2)$ is now being entertained with respect to *each* situation:

$$(<+4, -3, -2>, <+4, -3, -2>, <+4, -3, -2>) \tag{9.9}$$

[18] In private communication.

In this new state, the first component, $<+4, -3, -2>$, specifies my preferences with respect to s_1, the second component, which is exactly the same, specifies my preferences with respect to s_2, and so on. One might say that in this state I simultaneously accept three prescriptions that uniformly apply to all the three situations: one prescription *for* the action under consideration, with strength 4, and the other two *against* the action, with strengths 3 and 2, respectively.

But how is it possible to accept prescriptions that are mutually incompatible? How can I accept *both* that I ought to take the bike *and* that I ought not to do so? The answer is that the relevant ought-judgments are *pro tanto*: Each of them reflects just one relevant aspect of the case. In other words, they prescribe or forbid an action *insofar as* it has this-or-that feature. Thus, going by bike is prescribed insofar I originally prefer this action to be done in s_1, it is forbidden insofar I originally disprefer that it be done in s_2, in which I am in John's position, and it also is forbidden insofar I originally disprefer that it be done in s_3, in which I am in Mary's position.

Unlike oughts all-things-considered, *pro tanto* oughts are not overriding. The novelty of simultaneous extrapolation lies precisely in that it introduces *pro tanto* oughts. In other words, the novelty of this proposal is that it employs the universalizability requirement at an earlier stage than Hare himself: at the stage at which we do not yet commit ourselves, not even tentatively, to an overriding moral judgment all-told.

The remainder of the deliberation process goes as follows. In the state

$$(<+4, -3, -2>, <+4, -3, -2>, <+4, -3, -2>) \tag{9.10}$$

I have mutually conflicting preferences with respect to each situation $s_1 - s_3$. This intrapersonal preference conflict is then dealt with by straightforward balancing,

$$+4 - 3 - 2 = -1 \tag{9.11}$$

Consequently, I end up with the same preference all-told with regard to each situation:

$$(-1, -1, -1) \tag{9.12}$$

My overriding moral prescription all-things-considered, which is reached by the balancing of moral prescriptions *pro tanto*, is thus that I ought not to go to the office by bike, just as preference utilitarianism would have it: The stronger preference loses against the joined forces of the weaker preferences.

Above, we have noted that Conditional Reflection, if viewed as a conceptual truth, places serious hinders on preference change: Preferences in the input state are being entertained by the subject in accordance with Conditional Reflection. But then, if that principle is conceptually true, it seems impossible for the input preferences to ever be relinquished, as long as the subject retains full knowledge of what he would prefer if the roles were reversed. Now, none of them is yet relinquished in the step of simultaneous extrapolation, the one that takes the subject from $(+4, -3, -2)$ to $(<+4, -3, -2>, <+4, -3, -2>, <+4, -3, -2>)$. In that step,

the preferences are extended, but not given up. However, they do seem to be disappear in the final balancing step, as we move from $(<+4, -3, -2>, <+4, -3, -2>, <+4, -3, -2>)$ to the preferential state all-told $(-1, -1, -1)$. How can one deal with this problem?

The answer requires, I think, an appropriate interpretation of the balancing process. We go astray if we view it as a *reflective* formation of a new preference "all things considered", arrived at by consideration of the preferences we have previously acquired. Instead, this process should be seen more literally on the same model as addition of weights: The original preferences are added to each other, just as weights are added on a scale. This means that the original preferences are not given up in the final outcome, but remain present in the preference all-told. They are still there, as its different components.[19]

This reconstruction of the argument preserves Hare's conception of an ideal moral deliberation as a process in which

> [T]he interpersonal conflicts, however complex and however many persons are involved, will reduce themselves, given full knowledge of the preferences of others, to intrapersonal ones. (Hare 1981, p. 110)

However, the simultaneous extrapolation approach departs from Hare's moral theory at one crucial point: with respect to the overridingness issue. To be sure, Hare himself recognizes the possibility of overridable moral judgments, but these are according to him always *prima facie*.[20] They seem to hold 'at first sight', but may turn out to be invalid, in a particular case, upon further reflection. It is only in this sense that they can be overridden: They can be recognized as incorrect. Such overridable judgments are based on general "prima facie principles", which hold for the most part but admit of exceptions.[21] As far as I know, Hare never considered the possibility of moral judgments *pro tanto*, which retain their weight and validity even in those cases when they are being overridden (= outweighed) by other moral considerations.[22] Simultaneous extrapolation therefore requires that we go beyond Hare at this point.

While allowing for *pro tanto* oughts may be unproblematic, it is less clear what on this approach justifies the step from a preference I entertain with respect to some

[19] For pressing this point I am indebted to Mark Schroeder and Daniel Elstein. That Hare interprets balancing in this essentially 'non-reflective' way was suggested in Rabinowicz (1989). There, I contrast what I call "the data view" of preferences, according to which "the agent looks as his preferences [...] as *data*, as something that he can give weight to or discount when arriving at a decision" (ibid., p. 146), with "the driving-force view", which takes the agent's preferences at the time of decision to be jointly immediate determinants of choice. I suggest it is the latter view, which does not involve any intervening reflection step, that is more faithful to Hare.

[20] Unless they are what he calls 'inverted commas' moral judgments, "implying merely that a certain act is required in order to conform to the moral standards current in society" (Hare 1981, p. 58). 'Inverted commas' moral judgments are not genuinely moral, since they lack prescriptive force.

[21] Cf. Hare (1981), 59f.

[22] For a distinction between *pro tanto* and *prima facie*, as applied to reasons, see Kagan (1989, p. 17).

situation to its extrapolation – i.e. to a corresponding *pro tanto* moral prescription. The answer cannot simply be that I am trying to reach a moral judgment, which requires universality. It's certainly true that I am after a universal prescription *all-told*, but on what grounds do I first extrapolate each preference in my input state, i.e. frame universal prescriptions *pro tanto*? In response to this query, Elstein suggests that extrapolated preferences may be seen as judgments about moral *reasons*, "since each makes a contribution to the overall moral judgment, and it makes sense to say that the preference extrapolated from my preference in s_1 is still a consideration in favour of taking the bike (a reason) even after the overall judgment goes against" (private communication, Elstein, D., 2008). If this is how extrapolated preferences should be viewed, then the argument for extrapolation turns on the claim that moral reasons, just as moral judgments, are universalizable: A moral reason that applies to one situation, must also apply to every situation that is exactly similar, apart from the fact that the roles of the individuals have been reversed. So, if preferences I have with respect to s_1, s_2, or s_3 are all to be able to function as moral considerations in favour or against the universal moral judgment on what ought to be done in all these situations, each of them must itself be universalized, i.e. extrapolated. As Elstein puts it: "that argument involves the assumption that reasons are universalizable, which is not one that Hare discusses (as far as I remember). But it's a pretty natural companion to the view that moral judgments are universalizable, and may even be a corollary of that. [...] So, in brief, I think the simultaneous extrapolation can be motivated in a pretty Hare-friendly way by thinking about reasons" (ibid.).

9.5 Vendlerian Twist

Let me now turn to Zeno Vendler's comments on Hare's argument. It seems that, if Vendler is right, we might have all along been on a wild-goose chase. If he is right, the No-Conflict Problem is spurious. It should be dissolved rather than solved.

As we remember, the problem in question arises because the thought-experiments needed for Hare's argument concern purely *hypothetical* situations. When I ask myself what it would be like to be in someone else's shoes, and what preference I now have regarding that situation, I am supposed to consider a hypothetical state of affairs, which differs from the actual one in the role I occupy. My preference regarding what is to be done in such a hypothetical situation does not, on the face of it, conflict with my preference with respect to the actual situation, and this is what the No-Conflict Problem is all about.

The whole picture would change, if – as one might argue – the envisioned situation is not really distinct from the actual one. Suppose that what I consider in a thought-experiment is still the *actual* situation, but now viewed from *another perspective* – from the perspective of another person. If this is the case, then the preference I form regarding what is to be done in that situation does conflict with the preference I have when I view the very same situation from my own point of view. The resulting intrapersonal preference conflict can then be solved in the standard

way – by balancing. The action to be prescribed is the one that satisfies my conflicting preferences to the largest extent.

That imagining being exactly like someone else is in the actual situation does not take us to another possible world is a view that Vendler insists on:

> If I imagine being you, I do not imagine 'transporting' something into your body, or 'mixing' two entities. What I do is assume, as far as it is in the power of my imagination, the coherent set of experiences corresponding to your situation (your 'Humean self', as it were). But, as Hume pointed out, there is no specific experience of an 'I' in that totality. Nor is there one in mine. The 'I' as such has no content: it is the empty frame of consciousness indifferent to content. Consequently, by constructing in my fancy the image of what I would experience in your situation, I *ipso facto* represent your experiences. (Vendler 1988, p. 176)

> [I]n fancying being you or Castro, I do not touch the world: I merely switch perspectives on it. This is the reason, by the way, for my maintaining throughout this paper that imagining being in exactly the same qualitative conditions as another person is the same thing as imagining being that person. (ibid., p. 182)

> [T]here *seem* to be two different situations envisioned [...] Hare says this, 'Note that although the two situations are different, they differ only in what *individuals* occupy the two roles; their *universal* properties are all the same' (*MT* 111). 'No', I say, it is the same situation, with the same individuals; the only difference is which of them is I: in imagining being he, I imagine the same situation from a different perspective. (ibid., p. 178)

Vendler's position is very attractive. It does seem plausible to say that Hare's thought experiments do not target new situations, objectively speaking. Instead, they only effect a shift of subjective perspective. What 'moves' in such an experiment is not myself, the person I am, but only the "transcendental I", to use Vendler's Kantian terminology – a mere frame of consciousness that in principle can be filled with any content. Remember, that when I imagine being as John is now, I am not only supposed to take over his external circumstances but also his psychological make-up: his beliefs, emotions, desires, and so on.

Subjectively speaking, then, the situation changes when it is viewed from different perspectives. But, objectively, it still is the same situation. Therefore, when I form preferences that reflect the ones I ('the transcendental I') would have in different positions, all these preferences concern one and the same objective situation. As such, they can conflict with each other – the No-Conflict Problem is spurious.

Vendler himself thinks that this 'anti-metaphysical' move makes Hare's argument an easy sailing. However, it seems to me that the opposite may be the case. While the No-Conflict Problem disappears, we now get a new, more serious problem instead.[23] I still need to form preferences reflecting the ones 'I' would have in different positions; I need to entertain all these preferences together, from one and the same perspective, in order to balance them against each other. For this I have to rely on something like Conditional Reflection. But, and here comes the catch, that principle does not really seem to be applicable in contexts like this. Why not? Well, Conditional Reflection is an expression of *self-concern* – a fundamental attitude of caring for oneself that each well-integrated person is supposed to have. Self-concern applies not only to the actual situation; it also extends to hypothetical circumstances

[23] Cf. Section 7 in Rabinowicz and Strömberg (1996).

in which one might be placed. It manifests itself in the endorsement of the preferences one would have in hypothetical cases, just as Conditional Reflection has it.

Now, one might ask, who is it I am concerned about when I am concerned about *myself*? Is it 'the transcendental I' – a frame of consciousness that can be filled with an arbitrary content – or is it rather a definite *person*, the person I am? If it is the latter, as seems to me rather obvious (what do I care about a mere frame of consciousness?), then self-concern simply has no role to play in the radical thought-experiments of the kind Hare invites us to consider. In these experiments, what I envision is really being someone else, a different person. So it is not the question of imagining oneself – the very person one is – being placed in some hypothetical circumstances. But then, if self-concern does not extend to transcendental 'perspective shifts', such shifts remain outside the domain of application of Conditional Reflection. This means that Hare's argument cannot go through, contrary to what Vendler might have thought. The preferences belonging to different subjective perspectives concern the same objective situation, but, if Conditional Reflection is inapplicable, they need not all be reflected in a single perspective: They need not give rise to a co-existing set of preferences that are being entertained together, in one preference state. Consequently, they need not give rise to an intrapersonal conflict that can be solved by balancing.

A referee for this volume, Peter Dietsch, questions in his report my reasoning above:

> I do not understand why [...] a modified Conditional Reflection Principle [i.e. one that would be applicable to Hare's radical thought experiments, if Vendler was right in his anti-metaphysical move] is not available. Consider the following candidate, adapted to the terminology of Vendler's arguments: "Insofar as I fully know what I would prefer from the perspective of someone else in the actual situation, I must have the corresponding preference (same sign, same strength)." It is not clear to me in what sense the Conditional Reflection principle is wedded to a notion of self-concern [...] that would make this kind of application impossible.

It is a fair worry, but it seems to me that the modified Conditional Reflection is much too strong to be acceptable to someone like Hare – whether that principle is understood as a conceptual claim (which is Hare's own take on Conditional Reflection) or as a requirement of rationality. What the modified principle says, essentially, is that *mere empathy necessitates sympathy*: If I fully understand what someone in the actual situation desires, I must thereby come to have the corresponding desire. Surely, this view is foreign to Hare. As long as we are not yet in the business of making moral judgments, he would say, we aren't committed to sympathizing with our fellow beings, either on rational or on conceptual grounds. Mere empathy is not enough. There is no contradiction in the idea of a rational amoralist who is empathetic when it suits him, but who doesn't sympathize with his fellow beings.

To conclude: I am unsure whether Vendler is right in his 'anti-metaphysical' move. Anyway, to decide this matter would require an extended excursion into the area of modal metaphysics. *If* he is right, i.e. if Hare's thought experiments do not really take us to other possible worlds, then Hare's argument doesn't get off the ground, because Conditional Reflection is not meant to apply to transcendental

perspective-shifts. If he is wrong, on the other hand, or if the thought-experiments with role reversals could be given a less radical reading than the one Hare has in mind,[24] then such experiments would manage to take the subject beyond the actual situation to other, hypothetical situations in which he finds himself at the receiving end of the action under consideration. Then Conditional Reflection would apply, but we face the No-Conflict Problem. Its solution requires one's preferences be universalized, which can be implemented either by the preference revision approach or by the device of simultaneous preference extrapolation. In the former, a moral judgment all-told is arrived at directly, while in the latter it is only reached by the mediation of moral judgments *pro tanto*. Both approaches depart from Hare's own presentation of the process of moral deliberation, but the indirect approach seems to me preferable. For one thing, it is simpler and much closer to Hare's original formulation of the argument. It also is less question-begging. As we have seen, the preference revision proposal faces some serious problems. In particular, the defence of the principle of Minimal Change for preference revision and the choice of an appropriate measure of distance between preference states might turn out to lead to insurmountable difficulties.

References

Bratman, M. 1987. *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Gibbard, A. 1988. Hare's analysis of 'ought' and its implications. In *Hare and critics: essays on moral thinking*, eds. D. Seanor and N. Fotion, 57–72. Oxford: Clarendon.

Hare, R. M. 1963. *Freedom and reason*. Oxford: Oxford University Press.

Hare, R. M. 1981. *Moral thinking: its level, method and point*. Oxford: Clarendon.

Hare, R. M. 1987. Why moral language? In *Metaphysics & morality*, eds. P. Pettit, R. Sylvan, and J. Norman. Oxford: Basil Blackwell.

Hare, R. M. 1989. Reply to Ingmar Persson. *Theoria* 55: 171–177.

Harsanyi, J. C. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61: 434–435.

Harsanyi, J. C. 1977. Morality and the theory of rational behaviour. Social Research 44: 623–656. Reprinted in *Utilitarianism and beyond*, ed. A. Sen and B. Williams, 39–62. Cambridge: Cambridge University Press.

Gärdenfors, P. 1988, *Knowledge in Flux*. Cambridge, Mass.: The MIT Press.

Kagan, S. 1989. *The limits of morality*. Oxford: Clarendon.

Kant, I. 1785. *Grundlegung zur Metaphysik der Sitten*, 2nd ed., in H. J. Paton, *The Moral Law*, London: Hutchinson University Library.

Persson, I. 1989. Universalizability and the summing of desires. *Theoria* 55(3): 159–170.

Rabinowicz, W. 1989. Hare on prudence. *Theoria* 55: 145–151.

[24] An alternative would be to let the hypothetical situations the subject needs to consider in order to arrive at a moral judgment be similar to the actual situation in all the *relevant* respects, rather than in *all* respects, as Hare would have it. But this recourse to relevant similarities leads to its own problems: Deciding what aspects of the actual situation are morally relevant seems to be question-begging, if it is done at the *outset* of moral deliberation.

Rabinowicz, W. and Strömberg, B. 1996. What if I were in his shoes? On Hare's argument for preference utilitarianism. *Theoria* 62: 95–123.
Schueler, G. F. 1984. Some reasoning about preferences. *Ethics* 95: 78–80.
van Fraassen, B. 1984. Belief and the will. *Journal of Philosophy* 81: 235–256.
Vendler, Z. 1988. Changing places. *In Hare and critics: essays on moral thinking*, ed. D. Seanor and N. Fotion, 171–184. Oxford: Clarendon.

Chapter 10
The Ethics of *Nudge*[1]

Luc Bovens

Abstract In their recently published book *Nudge* (2008) Richard H. Thaler and Cass R. Sunstein (T&S) defend a position labelled as 'libertarian paternalism'. Their thinking appeals to both the right and the left of the political spectrum, as evidenced by the bedfellows they keep on either side of the Atlantic. In the US, they have advised Barack Obama, while, in the UK, they were welcomed with open arms by the David Cameron's camp (Chakrabortty 2008). I will consider the following questions. What is *Nudge*? How is it different from social advertisement? Does *Nudge* induce genuine preference change? Does *Nudge* build moral character? Is there a moral difference between the use of *Nudge* as opposed to subliminal images to reach policy objectives? And what are the moral constraints on *Nudge*?

10.1 The Paradigm Cases

I take *Cafeteria* and *Save More Tomorrow* to be paradigm cases of what constitutes a T&S-style *Nudge*, as these cases are repeatedly discussed in their writings (Sunstein and Thaler 2003, pp. 1159–1160 and pp. 1164–1166; T&S 2003, p. 175 and p. 177; T&S 2008, pp. 1–3 and pp. 112–115).

In *Cafeteria*, the goal is to induce students to choose a healthier diet. Studies show that individuals are prone to select items placed earlier and at eye level in a line of food items. So the school's management might try to affect students' diets by rearranging the display of the food items so as to make it more likely that the healthy food items are selected.

L. Bovens
London School of Economics and Political Science, Department of Philosophy, Logic and Scientific Method
e-mail: L.Bovens@LSE.ac.uk

[1] I am grateful for helpful comments to the audiences of the Models of Preference Change Workshop at the GAP6 and of the Choice group in the LSE; to the editors Till Gruene-Yanoff and Sven Ove Hansson; and to Foad Dizadji-Bahmani, Alice Obrecht, Adam Oliver, Esha Senchaudhuri, Peter Sozou, and Alex Voorhoeve.

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

In *Save More Tomorrow*, the goal is to make employees invest more in their pension fund savings. Employees are asked well ahead of time whether they are willing to commit next year's raise towards their pension funds. They are much more willing to agree to this than when they are asked to do so *after* they have received pay checks with the raises included. This exploits two psychological mechanisms. First, there is the *Endowment Effect*. People tend to find it much harder to part with something once they have it in hand than to forego it when they have never had it in hand. A description of this psychological mechanism can be found in David Hume's *Treatise*[2] and Adam Smith's *The Theory of Moral Sentiments.*[3] Second, people tend to find it much easier to show strength of will when it comes to future than to present costs and benefits. Think of Augustine's Prayer – *make me chaste, but not yet*. What explains the greater willingness to commit to a future loss of income to savings rather than a present loss is a combination of both of these psychological mechanisms.

Nudge is replete with examples that have a structure that is similar to these two cases. What these examples have in common is a manipulation of people's choices via the choice architecture, i.e. the way in which the choices are presented to them. This works in the following way. Choices are structured such that some psychological mechanism leads people toward options that are either thought to be in their own best interest or thought to be in society's best interest. In all cases of *Nudge*, if the choice situation had not been so structured, then people would be less prone to make the choice that is either in their own or in society's interest.

10.2 Social Advertisement

There is a more familiar type of intervention that the government employs to affect our behaviour. In social advertisement campaigns, we are made aware of the dangers of drug usage, the problem of domestic violence, the threat of AIDS, etc. How are such campaigns different from *Nudge*?

Social advertisement affects our choices by providing us with information or by affecting our emotions. Sometimes we learn things that we did not know before and change our behaviour. For example, an addict may change her drug habits after being informed of the death rate associated with cocaine usage. Other times there is no new information offered, but the situation is presented with such force that we change our behaviour. For example, pictures of domestic abuse may induce a wife beater to seek professional help.

T&S do discuss cases of framing in social advertisement (2008, pp. 180–182). For example, social advertisement that conveys the percentage of people who are registered as organ donors is more effective than if it were to convey the percentage

[2] 'Men generally fix their affections more on what they are possess'd of, than on what they never enjoyed (. . .)' (Hume 1978, Bk III, Part II, Section I; p. 482).

[3] 'To be deprived of that which we are possessed of, is a greater evil than to be disappointed of what we only have an expectation' (Smith 1968, Part II, Section II, Chapter II, p. 94).

of people who are not registered. The information provided is the same, but people are more likely to change their behaviour when the information is positively framed.[4] So social advertisement can be a form of *Nudge*, but not all social advertisement is *Nudge*. So what makes *Nudge* different?

10.3 Rationality and Autonomy

T&S write that their 'basic source of information' is 'the emerging science of choice, consisting of careful research by social scientists over the past 4 decades… [that] has raised serious questions about the rationality of many judgments and decisions that people make' (2008, p. 7). So one defining characteristic of *Nudge*, as opposed to social advertisement that does not qualify as *Nudge*, could be that some pattern of irrationality is being exploited. The psychological mechanisms that are exploited in *Cafeteria* and in *Save More Tomorrow* typically work better in the dark. If we tell students that the order of the food in the *Cafeteria* is rearranged for dietary purposes, then the intervention may be less successful. If we explain the endowment effect to employees, they may be less inclined to *Save More Tomorrow*. And even if we try to affect our own behaviour by means of these mechanisms, then our efforts will be most effective when our knowledge of having done so is latent (or when we simply are able to forget).

The following oft-cited example illustrates this well. People are prone to add an expensive car radio to their newly bought car. But if the car radio is not available on the day of the purchase and they are offered the very same car radio the very next day, then they would never dream of spending this kind of money on a car radio (Savage 1954, p. 103). Now once you point this out to them, they typically try to self-correct. They refrain from buying the expensive radio at the earlier point of time. Or, they may take this to be an argument for spending the money the next day – they remind themselves that they were perfectly happy to buy the radio on the day of purchase. It is not clear what direction they will take the argument, but at least, they will strive for less inconsistency in their actions.

There is something less than fully autonomous about the patterns of decision-making that *Nudge* taps into. When we are subject to the mechanisms that are studied in 'the science of choice', then we are not fully in control of our actions.[5]

[4] An alternative way of distinguishing *Nudge* from social advertisement is to stipulate that a *Nudge* must affect the actual choice situation. So a billboard with cancerous lungs is not a *Nudge*, but a pack of cigarettes with cancerous lungs is a *Nudge*. (I owe this suggestion to Alice Obrecht.) Or we could stipulate it as an additional condition on a *Nudge*. This would be in line with many of T&S's examples, but their example of social advertisement in support of organ donations that appeals to the framing effect would no longer qualify as a *Nudge*.

[5] We may of course use such patterns in an autonomous manner to steer our own agency – as when I rearrange the fridge myself when I am on a diet. But that does not make the action itself of picking the carrots placed in front an autonomous action. My agency is caused by processes that do not constitute reasons. In my quest for weight loss, I can autonomously set up a choice architecture that

When I am presented with full knowledge, then I tend to self-correct my agency. It seems that I was acting on a rule with which I cannot identify. What is so special about the first available item that I would favour it over later items? What is so special about having something in hand that would make it so much more valuable compared to the moment before I have it in hand? Clearly these are cases of not letting my actions be guided by principles that I can underwrite. And in as much, these actions are non-autonomous. Can they be said to be irrational? They can in so far as what is driving my action does not constitute a reason for my action – i.e. it is not a feature of the action that I endorse as a feature that makes the action desirable.

This brings us to the question with which we ended the last section. Why is at least some social advertisement different from *Nudge*? When social advertisement provides us with information that gives us a reason to change our behaviour then the intended effect is again fully autonomous decision-making. If it does not provide us with new information, but increases the saliency of certain reasons then the intended effect is again fully autonomous decision-making. And in this respect there is no reliance on the science of choice that raises questions about the rationality of our decision-making. Of course, these kinds of distinctions are much less clear in the real world. If a social advertiser frames the information in a particular manner that is known to have a greater impact on our decision-making – the more so if we are not made cognisant hereof – then we bring in elements of *Nudge* again. Reasons in support of (or against) the targeted agency and the causal mechanisms that raise (or diminish) the occurrence of the targeted agency mix together in social advertisement; in so far as social advertisement relies on the latter it has a bit of *Nudge* in it.

10.4 What Type of Agency Does *Nudge* Aim to Correct?

I will distinguish between six types of agency that can be made the subject of a *Nudge*.

(i) ***Ignorance***. If the government sets a default for retirement plans, they may do so for the same reason that a medical doctor might recommend a treatment. We typically have little knowledge of the matter at hand. We have a clear goal, viz. to be well off in old age or to recover from an ailment. But the route to this goal requires special expertise, which we lack. So it is lack of knowledge that hampers us in laying out the steps towards realising our goals.

(ii) ***Inertia***. It may be the case that we do have sufficient knowledge, but somehow inertia gets the best of us. We are absorbed in our daily activities and simply put off filling out the forms until we forget. In this case a default option kicking in for the lazy or forgetful may be welcome.

---

will induce me to act non-autonomously. Consider the following analogy. If I am prone to squeeze a stone in my pocket for good luck, then it may be fully rational to do so when I need to work up the self-confidence in, say, an interview situation. But this does not make the action of squeezing the stone itself rational (cf. Bovens 1995, p. 824).

(iii) ***Akrasia***. Consider our paradigm cases again. Typically we know quite well that our consumption of cream puffs is not conducive towards overall health. We know quite well that we are putting too little money into our retirement funds. What stands in the way is weakness of the will (*akrasia*). We are weak-willed in choosing the proper steps towards our long-term goals. By structuring the choice situation it becomes easier to correct for such weak-willed actions, because temptation will have less of a pull on us.

(iv) ***Queasiness***. In *post-mortem* organ donations, the culprit is not lack of knowledge or weakness of the will. Many of us have no objection to becoming organ donors. It may be inertia, but it could also be queasiness, which prevents us from becoming donors. We are perfectly fine with our organs being used *post-mortem* for transplantation, but we do not want to entertain such matters in decision-making. There is an emotional cost in making the decision of becoming a *post-mortem* organ donor.

(v) ***Exception***. Suppose that particular choices by people with a particular profile tend to engender feelings of regret, whereas alternative choices tend to induce greater *ex post* satisfaction. Let us suppose that there is ample evidence for this in empirical research. For example, we could think of sex change surgeries, abortions, divorces, teenage sex, or what have you. It may well be the case that it holds true as a statistical claim that people who choose some such options typically experience feelings of regret afterwards. Of course there is a reference class problem. It may be the case that for the subgroup to which I belong, suitably defined, this tendency is false. For example, although most transgendered people experience regret after a sex change surgery, the subgroup of transgendered people of which I am a member (say, female, engaged in a relationship with an accepting partner, ...) does not. But suppose that for the narrowest social group for whom we can obtain meaningful results and of which I am a member, this tendency holds. Then I could still claim that, though most people display such feelings of regrets, I, for one, am confident that I will not. And I may be correct in my claim.

(vi) ***Social Benefits***. It may well be the case that a particular individual choice is not socially beneficial. There are many such examples. I may see no benefit whatsoever in giving to charity. In a Tragedy of the Commons, society may be better off if I decide to refrain from, say, adding a fishing boat to over-fished waters or drilling for oil in an oil well that is already quite depleted. But unless I am being compensated or find alternative employment, I may be worse off. In a standard Prisoner's Dilemma, society would have been better off if we had all cooperated, but I would have been worse off if I had cooperated rather than defected. In all such cases, a bit of *Nudge* might be meaningful to realise a socially beneficial outcome, but it may well be at the cost of my own welfare.

In the real world, these distinctions are less pronounced and there are many grey cases. Some cases of *Inertia* may be instances of convenient forgetfulness that are not altogether different from *Akrasia*. *Ignorance* may be intentional because we wish to forego making hard decisions and in this respect such cases may not be altogether different from *Inertia* and *Queasiness*. And I may simply adjust my overall

preferences when I come to learn about the statistical evidence or social benefits and then the only thing that would block my ability to act in a particular case is *Akrasia*. But the existence of these mixed cases does not invalidate the exercise of distinguishing between ideal cases, which will prove useful when thinking about preference change and the moral permissibility of *Nudge*.

10.5 Preference Change[6]

Let us start with *Exception* and *Social Benefits*. In these cases, I am being *Nudged* in the direction of agency which I do not believe to be in my interest. For instance, suppose I believe upon reflection that there is no need to increase my pension savings, but the *Save-More-Tomorrow Nudge* did induce me to do so. What can we say about my preferences when I decide to invest my future raise into my pension fund whereas I would not have done so after the raise had been in place? Did I undergo a preference change?

In one respect the answer is yes – I revealed my preference through my choice. What has changed is that I have a preference for dedicating a greater percentage of my income to my pension fund here and now, which I never had before. In another respect the answer is no. Have I become a more frugal person? Not really. In a way my action is aberrant. It is not well integrated with my overall preference structure – i.e. with my conception of the good, with what I take to be good for me all things considered. I am like the fox and the sour grapes. The fox loses his appetite[7] for the grapes that he cannot reach. Even if he does not want these grapes anymore, he remains the kind of fox who likes juicy summer fruits in general. So his preference over the token action of eating these very grapes does not cohere well with his preferences over the type of actions of eating juicy summer fruits. Similarly, my preference for the token action of dedicating a greater percentage of my income to my pension fund does not cohere well with my preference for actions that are non-frugal in character. It is no different than signing the lease for a timeshare in the Virgin Isles with a clever salesperson in charge – in some respect, I do want it, but in another respect, I do not, since it does not fit in with my overall preference structure. We choose on the background of a fragmented self. In answering the question whether we do or do not want to buy into the *Save More Tomorrow* scheme, whether the fox does or does not want the grapes, or whether we do or do not want the timeshare, a gloss is needed – in some respect, yes, in another respect no.

Of course coherence may be regained by making changes in my preference structure at large. The fox may well turn away from juicy summer fruits in general after

[6] This section builds on Bovens (1992).

[7] This is the preference change interpretation of the fable, which we find in Elster (1983, p. 123). This interpretation differs from a case of self-deception in which the fox would turn away and say that *these* grapes are too sour, are unripe, or what have you. This would be a case of a belief change induced by considerations of feasibility rather than by evidence for the proposition at hand.

a few bad experiences with fruit that is beyond his reach. Similarly, I may acquire a taste for frugal actions after a few *Nudges* in this direction. There are various mechanisms that may bring this about. I may come to appreciate such actions through discovering hitherto unknown attractive features. I may become habituated in Aristotelian style – my feelings may simply shift by repeatedly acting frugally or eating healthy foods. I may come to self-identify as a person who acts frugally or eats healthy foods on grounds of cognitive dissonance. All such mechanisms could be successful in bringing about preference shifts over action types and then my newly acquired preference is genuine.

What can we say about *Nudge* in cases of *Ignorance*, *Inertia*, *Akrasia*, of *Queasiness*? In these cases *Nudge* steers us in the direction of what we consider to be in line with our overall preference structure. Our initial preferences over action tokens do not cohere well with our overall preferences. So now we are *Nudged* in a direction that restores coherence between our actions and our overall preference structure. Is this a genuine preference change?

There is another lesson to be learned from the fox. Suppose that the fox tells us upon sincere introspection that he does not want the grapes anymore. Suppose that he even changes his overall preference structure and turns up his nose for juicy summer fruits in general. But now suppose that we lower the vine for the fox. Would he reconsider? Suppose that he would, maybe not immediately, but after encountering a few low-hanging bunches of grapes, he would be at it eating grapes again. Then we would need to qualify the claim that the fox changed his preferences. Again, we would add a gloss – we would say that his preference change was too short-lived to qualify as a genuine preference change. Similarly we would be hesitant to say that the *Nudge* made us prefer to dedicate a greater percentage of our income to our pension fund, when we would decide differently without the aid of some clever choice architecture. Even though we may have an overall preference to be more frugal, we can hardly be said to genuinely prefer to dedicate a greater percentage of our income towards our pension fund next year when this type of action is not resilient without the aid of *Nudges*.

As before, it may be the case that a few *Nudges* provide me with a taste for frugal actions and that it becomes easier for me to act in accordance with my overall preference structure, also in the absence of clever choice architectures. Then the non-resilience objection drops out and the preference change becomes genuine.

Let me sum up by making the point by means of *Cafeteria*. Suppose that a *Nudge* succeeds in making me take the healthy snack. Did it then induce a preference in me for the healthy snack? In some respect, yes – I revealed my preference through my choice. But this may need a gloss. Suppose that I am actually the person who values the life style of the glutton. Then in another respect, I do not genuinely prefer the healthy snack. This case maps onto the cases of *Exception* and *Social Benefits*. Or suppose that I do prefer a healthy life style, but I continue to take ice-cream under non-*Nudge* condition. Then you would look at me strangely if I were to proclaim that I genuinely prefer the healthy snack when I am placed under *Nudge* conditions. You would point out to me that I just took the healthy snack because of the choice architecture – I do not genuinely prefer the healthy snack. This case maps onto

the cases of *Ignorance*, *Inertia*, *Akrasia*, and *Queasiness*. However, if adjustments to the overall preference structure are made in *Exception* and *Social Benefits* or resilience is gained in *Ignorance*,. . . , *Queasiness*, then the preference change becomes a genuine preference change and we no longer need glosses.[8]

10.6 Does *Nudge* Build Moral Character?

We continue with cases of *Ignorance*, *Inertia*, *Akrasia*, and *Queasiness*. It is a lack of self-control that blocks us from acting in accordance with our overall preference structure. So when we are *Nudged* in the direction of actions that we take to be in our interest all things considered, does this build moral character? Does it increase our capacity for self-control?

The folk singer Karen Dalton once said that she sang softly because she wanted people to listen to her. This strikes us as paradoxical. Certainly people are more likely to listen when you raise your voice. Indeed, this is the expectation of the short-term effect. But the long-term effect may be precisely reversed. Think of a grade school teacher who is prone to raise her voice. This may be effective in the short term, but she may have to raise her voice more and more and the overall effect may be that more children would have listened to her had she never raised her voice to begin with. Similarly, there is research showing that the death penalty has a deterrence effect in that the rate of pardon by the governor correlates with the rate of violent crime in subsequent years (Gittings and Mocan 2003). This is consistent with the brutalisation effect – capital punishment may contribute to a more violent culture and may increase violent crime in the long run.

Now it may be the case that repeated *Nudging* in public health and pension funds may have short-term positive effects at best. *Nudging* may not create sustainable effects on people's behaviour for the long-term; as time goes on, the level of *Nudging* required to retain this effect may increase. Just as Karen Dalton did not want to raise her voice, knowing full well that some people would zone out, we should not

[8] I would like to flag the following nagging concern. Some people may object that if we regain coherence through changes to our overall preference structure (in *Exception* and *Social Benefits*) or to our particular preferences under non-*Nudge* conditions (in *Ignorance*, . . . , *Queasiness*) through mechanisms such as habituation, cognitive dissonance, . . . , then this is worrisome because of the broad scope of non-autonomous preference change. (This objection was raised by Jason Alexander and Alice Obrecht.) In "Sour Grapes and Character Planning" (1992), I argued in response to Jon Elster (1983, pp. 24–25) that it is not the lack of autonomy of the fox's preference change, but rather the lack of coherence between his adjusted preference and his overall preference structure. And this is what distinguishes sour grapes from character planning. In character planning, there is an adjustment of the particular preference as well as an adjustment of our overall preference structure so that coherence is restored. The lack of autonomy in preference change is unproblematic – our preferences may be fully rational even though we did not autonomously acquire them. I am comfortable repeating this line when it comes to *Ignorance*,. . . , *Queasiness*, but slightly nervous when it comes to *Exception* and *Social Benefits*. But elucidating this difference – if indeed there is such a difference – will require more reflection.

be lured by the short-term success of *Nudging* either. To warrant long-term success, we should let people make their own decisions while providing minimal aid. My point is that short-term success of *Nudge* may be consistent with long-term failure. The long-term effect of *Nudge* may be infantilisation, i.e. decreased responsibility in matters regarding one's own welfare.

But of course things ain't necessarily so. Cognitive dissonance, habituation, acquiring a taste for the good-making features in *Nudged* actions may bring about long-term preference change as well. More people may come to adjust their overall dietary habits (and not only in the *Cafeteria* setting) or become more prudent in general (and not only in the *Save More Tomorrow* scheme.) This brings us back to the question of whether *Nudge* induces genuine preference change. When we come to acquire a taste for the *Nudged* actions, then the effects will be more broad-ranging and long-lasting.

At the end of the day, different people will be affected in different ways and it is an empirical question whether there does exist something like the infantilisation effect, just like it is an empirical question whether there exist something like the brutalisation effect of capital punishment. My only aim here is to point out that, just as a study of the (short-term) deterrence effects of capital punishment by means of time-series analysis is not the last word, a study of the (short-term) success of a particular *Nudge* is not the last word either. Granted, brutalisation and infantilisation effects are difficult to study through empirical testing. It does not suffice to do cross-population studies and to point to the correlation between capital punishment and the number of executions on the one hand and the rate of violent crime on the other hand, since the causal direction is unclear. Counter to the brutalisation effect, it may well be the case that high rate of violent crimes is the cause of the institution and the prevalence of capital punishment, rather than vice versa. The same problem would occur if we were to find a correlation between some measure of responsibility and paternalistic policies. Counter to the infantilisation effect, it may well be the case that the low measure of responsibility is the cause of the institution and the prevalence of paternalistic policies rather than vice versa.

10.7 Who Is *Nudging*?

It matters a great deal who is doing the *Nudging*. Let us start with a case in which I set up a *Nudge* to constrain my own behaviour. This is an example of *sophisticated choice* (McClennen 1990, p. 12). I may force myself to decide on increased pension-fund contributions earlier rather than later because I know that it is my only hope to commit a reasonable amount. I don't think that there is much to object here. Now some strong-willed people consider it to be wrong for me to decide, say, not to bring any liquor in the home. They seem to believe that we should educate ourselves so as to become *resolute choosers*, who are able to commit just as much to their pension funds after receiving a raise as before receiving it and who are able to drink just as little, whether there is liquor in the home or not. But let us bracket such ideals of

perfectionism. As long as we are self-legislating, there seems to be little to object to in engaging a *Nudge*.

But now let us go one step further. Suppose that I choose a nudging partner and consciously or unconsciously take his or her nudging to be a good-making feature. Or suppose that I choose to work in a self-professed paternalistic company. In either case, it seems that I have little to object to when the fridge or the line with food items is carefully arranged so that I am more likely to take the healthy options.

But this brings us to the actual concern with *Nudge* as a social policy instrument. What if a majority elects a government with a nanny-state platform? Do they have a democratic mandate to *Nudge*? What about the minority who does not want the government to interfere with their preference formation? We will return to this question below.

10.8 Transparency

Without going into the empirical details, let us suppose that the use of subliminal images could actually bring about preference change. Now typically the use of such devices makes people extremely nervous. Suppose that the government starts a public health campaign to reduce obesity. Let there be a social group with problematic dietary habits. Research shows that there is a high density of viewers from this social group for a particular TV programme. So we decide to splice pictures of happy carrot-eaters into this programme as subliminal images. T&S object to this practice because of the lack of transparency (2008, pp. 244–245). But then suppose that the government simply announces that it will combat social problems by means of subliminal images. T&S object that also this would not suffice, because 'manipulation of this kind is objectionable precisely because it is invisible and thus impossible to monitor' (2008, p. 246).

So how is this any different from *Cafeteria*? We need to make a distinction between *type interference transparency* and *token interference transparency*. It is one thing for the government to say that they will be using certain types of psychological mechanisms to solve social problems. This is type interference transparency – the government is transparent about how it will try to interfere with our agency. But then there is no difference between *Nudges* and subliminal images – the government can announce that it will *Nudge* and that it will use subliminal images. Yet T&S support the former and object to the latter. So type interference transparency is not enough.

I take it that T&S also wish to have token interference transparency. How does *Nudge* differ from subliminal images? Being exposed to a particular image at a particular time is a token interference by means of a subliminal image. When we are affected by such a token interference, there is no way that we could notice (blocking the use of special equipment). But if we are being *Nudged*, it is possible to recognise here and now that the food is arranged in a particular manner, that the pension savings forms are sent early to facilitate saving etc. So in a *Nudge*, it is possible to recognise each token interference.

So does this mean we need to put up a billboard next to the food line stating: "Research shows that people are more prone to take food items displayed earlier rather than further down the line. Many of our customers are trying to lose weight but find it difficult to do so. To help them, we have arranged the snacks in the food line with healthier items displayed earlier so that they are more likely to choose these items." The problem is that these techniques do work best in the dark. So the more *actual* token interference transparency we demand, the less effective these techniques are. But it may just be sufficient that there is *in principle* token interference transparency. A watchful person would be able to identify the intention of the choice architecture and she could blow the whistle if she judges that the government is overstepping its mandate. This in principle token interference transparency is not possible for subliminal images. In giving the government a mandate to use subliminal images we would be signing a blank check and could only hope that they will not be abusing their power and splice in ads for the incumbent in the next election.

In summary, subliminal images are deemed impermissible because they do not satisfy *in principle token interference transparency*, whereas T&S-style *Nudges* do pass this requirement. But then are we not confident that there are some watchdogs with sophisticated equipment keeping an eye on the government? Certainly, but I think that we find it important that also *we ourselves* could decide to become watchful and unmask any manipulation. In the democratic process we may give the government a mandate to engage in certain types of *Nudges*. But then we wish to respect the right of minorities who do not appreciate this type of manipulation. To safeguard their interests, we stipulate that every *Nudge* should be such that it is in principle possible for everyone who is watchful to unmask the manipulation.

10.9 The Moral Permissibility of *Nudge*

I have pointed to a number of issues that are relevant when we judge the permissibility of a particular *Nudge*.

First, it is less worrisome when the *Nudge* brings our agency in line with our overall preferences, as in *Ignorance, Inertia, Akrasia and Queasiness* than when the projected agency is not in line with our overall preferences, as in *Exception* and *Social Benefits*.

Second, it is less desirable when a *Nudge* is local and leaves us with a fragmented self. We become incomprehensible to ourselves – why did we not act in line with our overall preferences or why is this kind of agency not resilient under non-*Nudge* conditions? We can avoid such a fragmented self if our *Nudging* brings about change in our general preference structure or change in our agency that continues to hold under non-*Nudge* conditions. But there is a tension here. Some will undoubtedly be even more worried if *Nudge* brings about massive changes in our preferences through psychological mechanisms such as habituation, cognitive dissonance etc. Fragmentation avoidance comes at the cost of even more non-autonomous preference change. This may be worrisome in cases like *Exception* and *Social Benefits* (see footnote 8).

Third, *Nudge* is less desirable when it creates a people who have become incapable of taking their lives in their own hands and to make autonomous changes in their agency to make it fit in with their overall preference structure. Such long-term infantilisation effects are difficult to assess empirically but it is nonetheless a concern that does not go away. Adam Smith (Part VI, Section III; pp. 143–145) thought that adversity was the best school to develop the respectable virtue of self-command. The cost of *Nudge* may be that we forego the chance to gain the virtue of self-command.

Fourth, the less control we retain over being *Nudged*, the more problematic it is. If we choose to put ourselves into a situation that is rich with *Nudges*, then we have little to complain about. But does this type of consent extend to a democratic mandate to the government to be *Nudged*? I have argued that *Nudges* must be transparent in principle at the level of each token *Nudge*, in order to ensure that everyone can unmask the manipulation if they wish to do so. This protects the rights of the minorities who do not wish to be so manipulated and it keeps a check on the government.

There are many other factors that enter into the permissibility of *Nudge*. Let me just flag a few. Advertisement for products that do not increase welfare may use all kinds of *Nudge* style techniques and the government may be fighting a losing battle against, say, obesity, if it cannot access the same arsenal of techniques. Furthermore, governments commonly set up quasi-markets to increase efficiency in the provision of public goods. Citizens are bombarded by technical information from competing providers. Securing health insurance should not be as complicated as choosing a cell-phone. If the government institutes such quasi-markets then it also has the responsibility to navigate people through them which may involve more or less gentle *Nudging*. Finally, the more urgent the problem that a *Nudge* is trying to tackle, the less it meets with qualms. Instituting *Save for Tomorrow* may be more acceptable in the US than in South East Asia, considering differential saving rates. Instituting *Cafeteria* may be more acceptable in Chicago than in Paris, considering differential obesity rates. And, no doubt, in assessing the permissibility of *particular Nudges*, many more considerations that are idiosyncratic to the case at hand will emerge and each case will need to be assessed on its own merits.

References

Bovens, L. 1992. Sour Grapes and Character Planning. *Journal of Philosophy* 84: 57–78.

Bovens, L. 1995. The Intentional Acquisition of Mental States. *Philosophy and Phenomenological Research* 55: 821–840.

Chakrabortty, A. 2008. From Obama to Cameron, Why Do so Many Politicians Want a Piece of Richard Thaler? *Guardian* July 12.

Elster, J. 1983. *Sour Grapes*. Cambridge: Cambridge University Press.

Gittings, R. K. and N. Mocan. 2003. Getting Off Death Row: Commuted Sentences and the Deterrent Effect of Capital Punishment. *Journal of Law and Economics* 46: 453–478.

Hume, D. 1978. *A Treatise of Human Nature*. (2nd Ed.) Edited by L.A. Selby-Bigge. Oxford: Clarendon.

McClennen, E. F. 1990 *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.
Savage, J. 1954. *Foundations of Statistics*. New York: Wiley.
Smith, A. 1968. *The Essential Adam Smith*. Edited by R. L. Heilbroner. New York: Norton.
Sunstein, C. R. and R. H. Thaler. 2003. Libertarian Paternalism Is Not an Oxymoron. *University of Chicago Law Review* 70: 1159–1202.
Thaler, R. H. and C. R. Sunstein. 2003. Libertarian Paternalism. *American Economic Review* 93: 175–179.
Thaler, R. H. and C. R. Sunstein. 2008. *Nudge*. London: Yale University Press.

Chapter 11
Preference Kinematics

Richard Bradley*

Abstract Preferences, like beliefs, can and do change as a result of interaction with the environment and/or internal processes. This paper presents a kinematical model in which preference changes are explicated and motivated in terms of changes to the agent's quantitative degrees of belief and/or desire across some particular partition of prospects. Several basic types of such changes are identified and consistency conditions for them specified. Finally, the model is used to explain the possibility of preference loss and gain.

11.1 Introduction

Preferences, like other attitudes, can and often do change as a result of interaction with the environment – in response to observations, experimentation and verbal testimony, for instance – as well as a result of internal processes, both cognitive and biological – in response to deliberation or physical maturation, for instance. An understanding of how preferences change, or should change, as result of these processes is critical to a range of questions, both empirical and normative. For instance:

1. The normative problem of how to evaluate different potential social arrangements or institutions in terms of the preferences of the individuals affected by them cannot be adequately addressed without an understanding of the effect that the emergence and persistence of the arrangements will have on these individual's preferences.
2. Practical deliberation, especially as conducted by a group of agents, can hardly be described, let alone subjected to normative analysis, without some way of modelling the manner in which agents influence one another's preferences and beliefs.

R. Bradley
Department of Philosophy, Logic and Scientific Method, London School of Economics, Houghton Street, London WC2A 2AE
e-mail: r.bradley@lse.ac.uk

* I am indebted to Alex Voorhoeve, Christian List and Franz Dietrich for their help with the ideas contained in this paper.

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

3. Empirical work on processes such as socialisation and the spread of culture will remain disconnected from theories of rational, autonomous agency unless it is possible to model these effects at the level of agents' attitudes.

Despite this, work on preference revision is rather heterogenous and, in some fields, quite sparse. Until the recent growth of behavioural and evolutionary economics it was largely a neglected question in neoclassical economics whose models typically treated preferences as both exogenously given and stable.[1] To the extent that Robert Pollak [16] remarked that "those who favour incorporating taste formation and change into economic analysis fall into two groups whose intersection is almost empty", referring on the one hand to 'Radical' economists working on the ideological and institutional determinants of preferences and on the other to economists like himself interested in the dynamics of household consumption. The main exception is Gary Becker [1], though his important contribution to modelling preference change has been somewhat obscured by the reductionist nature of his account and his insistence that fundamental preferences or 'tastes' are invariant across both times and persons.

In philosophy the situation is not dissimilar, with surprisingly little attention being given to rational preference or value revision compared to, say, the wealth of work on belief revision. The most prominent exceptions are Hansson (see [10, 11]), whose approach has many affinities with the one taken in this paper, and the recent work by van Benthem and Liu and others (see [2, 14]). In philosophy too the view that fundamental preferences are invariant has a long history. It was essential to classical Utilitarianism, for instance, that preferences for pleasure over pain should be both universal and stable: this is what made quantities of pleasure the right currency for moral accounting. But the idea that preferences and other evaluative judgements can be derived from fundamental desires or values extends far beyond this tradition.

It is not my intention to assess the truth of this hypothesis: in all likelihood it is not subject to any direct empirical test and has to be assessed on broader methodological grounds as well.[2] Rather, I want to sketch out a framework in which rational preference change can be both described and evaluated without recourse to the postulate of fundamental preferences. The framework is properly best viewed as extension of the familiar Bayesian theories of belief revision to preferences and, in particular, Richard Jeffrey's kinematical version (see [13]). A similar quantitative representation of the states of mind of agents is employed and similar conditioning rules described. Like Bayesian conditioning, the models will admit of both normative and descriptive interpretations and results relevant to both will be presented.

I will proceed as follows. In the first section, some examples of preference change will be aired and the basic framework for representing them laid out. In the second a model of rational revision in response to experience will be presented, for the case in which the effects of the latter can be localised to a partition of the space of prospects. In the succeeding two sections rational taste-driven and belief-driven

[1] On this recent work see Bowles [1].

[2] See Bradley [8] for an examination of this view.

preference change will be analysed using this model. In the final section, we consider the question of incomplete preferences and the possibility of expansions or contractions of the domain of the preference relation.

11.2 Representing Preference Change

11.2.1 Types of Preference Change

Let us start by considering some examples of types of drivers or causes of preference change. The list is not intended to be exhaustive; nor is it necessarily the case that the items on it belong to mutually exclusive categories. It is simply intended as a starting point to help fix the scope of our discussion.

1. Re-evaluation of preferences over options in the light of new information about the state of the world, e.g. when you change your attitude to going to the beach when you learn that it is likely to rain.
2. Change in attitude to a possible event in the light of information about someone's intentions or actions, e.g. when learning that your neighbour is planning a barbecue makes you prefer that it rains that day.
3. Adapting your preferences for outcomes as their realisation becomes more or less probable e.g. when you lose interest in a sporting competition when it becomes clear when you have no hope of winning it.
4. Conditioning or cultivation of taste by habituation, e.g. weaning infants onto cow's milk or acquiring a taste for olives.
5. Discovering the value of things, e.g. when you learn that relationships require discretion as well as honesty, or that red wine is best drunk with cheese.
6. Changing preferences for an activity as the amount of anticipated pleasure deriving from it changes with increasing skill, e.g. learning how to play the piano, mastering the crossword puzzle.
7. Changes induced by discussion or persuasion, e.g. you change your attitude to a summer holiday in Italy when your wife reminds you that she hates crowds.
8. Preferences formed by inquiry or deliberation, e.g. when you develop a preference for one hotel over another by comparing the reviews on each or visiting each in turn.
9. Creation of new preferences, e.g. new goods come onto the market, you meet someone for the first time, or you are told about a restaurant that you had not heard of before.

Cases 1, 2 and 3 are clearly instances of preference change being driven by receipt of information or by a belief change. The first is the one most easily handled within traditional rational choice theory because it can be explained by a change in the expected utilities of the options induced by conditionalising on the new information. The second case is much more difficult to model in this framework, because here the preferences that change are those that are directed at the state of the world, rather than at actions. As we shall see, however, such cases can be elegantly dealt

with in the framework of Jeffrey decision theory. Case 3 differs from the other two in that the belief change causes the preference change in question without being a reason for it: the fact that I am unlikely to win a competition is no reason to regard winning it as less valuable than winning at one in which my prospects are good.

Cases 4, 5 and 6 look more like instances of taste or desire change rather than belief change.[3] In habituation cases repeated experience of something leads to a re-evaluation of it (typically unconsciously) despite the fact that any informational gains are made only in the early repetitions. One can grow tired of a foodstuff, for instance, not because of anything one learns about it, but simply because of the jading of one's palate. In cases of value discovery, some kind of learning is involved, but it seems to be of a different nature to that involved in the improvement of belief. When one learns that a particular wine is a good companion to a particular cheese (perhaps contrary to prior expectations), one does of course learn something about the two products. But what one learns about them is how they stand in relation to one's tastes; a discovery that must give rise to an improved evaluation of the products in combination, before it gives rise to a new (and improved) belief about them. Cases of skill change seem to have aspects in common with both habituation and value learning. What is special about them, however, is that the value is created rather than learnt or acquired; it is the mastery of the skill that makes the activity pleasurable.

Can all cases of preference change be explained in terms of changes in information, or more generally belief, or changes in tastes, or combinations of the two? It seems to me that a large number of cases can be. Preference change as a result of deliberation or discussion, for instance, typically involves both since the opinions of others (and indeed the deliberative process itself) can have both an affective or cognitive impact. More controversially, I will argue that some instances in which an agent forms a preference judgement between two prospects for the first time, or withdraws her previous judgement without replacing it with a new one, can be explained in these terms. Case 8 is an example of this.

But there are limits to this simple-minded explanatory strategy. On the face of it, for instance, it will not work in Case 9 where new preferences are created when the agent becomes aware of a possibility for the first time. Cases like this cannot be captured in the models that will be presented here, essentially because these models assume a given space of prospects. This is no doubt an important weakness, but not I think one that is insurmountable.

11.2.2 Framework

States of Mind In the approach taken here we think of an agent's preferences for prospects as being explained or rationalised, depending on whether the model is employed descriptively or normatively, by the agent's state of mind. An agent's state of mind is given by her degrees of belief and desire for some set of prospects. In the

[3] And for this reason receives a good deal of attention in Becker [18].

Bolker-Jeffrey framework that is adopted here, the set of prospects $\Omega = \{A, B, ...\}$ is assumed to form a Boolean algebra containing the unit, T, but with the zero, F, removed.[4] The join of any two prospects X and Y will be denoted by $X \vee Y$, their meet by XY and the complement of X by $\neg X$. In this paper we assume that the set of prospects is static and that what changes is the agent's attitudes to them.

The agent's preferences are modelled as a two-place relation on the set of prospects Ω, which we will require to be transitive, but not necessarily complete.[5] Prior preferences will be denoted by $\succeq$ and posterior ones (i.e. those after some attitude changing experience) by $\succeq^*$. The relations of strict preference, $\succ$, and indifference, $\approx$, are related to $\succeq$ in the usual way. Informally we may think of the agent's preferences, and their evolution over time, as providing the observations that require explanation or rationalisation in terms of changes to beliefs and desires.

The state of mind of an opinionated agent will be represented by a pair of real-valued functions $\langle p, v\rangle$, where p is a probability measure on Ω of her degrees of belief and v a normalised desirability function on $\Omega - \{F\}$ measuring her degrees of desire and satisfying, for all prospects X and Y such that $XY = F$:

Axiom 1 *(Normality)* $v(T) = 0$

Axiom 2 *(Averaging)* $v(X \vee Y) = \frac{v(X).p(X)+v(Y).p(Y)}{p(X)+p(Y)}$

The functions p and v also provide a basis for a representation of the agent's conditional degrees of belief and desire – her degrees of belief and desire under the hypothesis that some or other condition holds. This is achieved via the definitions:

Definition 3 (Conditional Probability) *If* $p(A) > 0$*, then:*

$$p(X|A) := \frac{p(XA)}{p(A)}$$

Definition 4 (Conditional Desirability) *If* $p(AX) > 0$*, then:*

$$v(X|A) := v(XA) - v(A)$$

A notable feature of this representation is that an agent's attitude to any prospect can be expressed as a weighted average of her attitudes to the various ways in which that prospect could be realised. Thus if $\{A_i\}$ is a partition:

$$p(X) = \sum_i p(XA_i) = \sum_i p(X|A_i).p(A_i) \tag{11.1}$$

$$v(X) = \sum_i v(XA_i).p(A_i|X) = \sum_i (v(X|A_i) + v(A_i)).p(A_i|X) \tag{11.2}$$

[4] See Jeffrey [12].

[5] We often speak of preferences for properties of objects, e.g. for honest shopkeepers over dishonest ones. The relation between property preferences and prospect preferences is an interesting one and no doubt a richer theory of preference change should reflect them. For discussions of this issue see [15] and [9].

An opinionated state of mind is clearly an extreme case and in this paper we will not assume that agents have formed judgements about all prospects. We may nonetheless use the same framework by representing the state of mind of a less than fully opinionated agent by a set of pairs of probability and desirability functions: intuitively the set of opinionated states of mind consistent with what the agent actually believes and desires.

Explanation and Justification Models of rational agents of the kind adopted here can be used for both descriptive and normative purposes to either explain (and sometimes predict) how agents will behave or to rationalise or justify choices. In particular, an opinionated state of mind $\langle p, v \rangle$ explains or justifies an agent's preferences whenever it is the case that, for all X, Y in the domain of $\succeq$:

$$X \succeq Y \Rightarrow v(X) \geq v(Y)$$

The more information we hold about an agent's preferences, the more constraints they place on the class of opinionated states of mind that explain them. Under certain assumptions about the preference relation $\succeq$ – completeness and Bolker's averaging and impartiality conditions (see Bolker [3]) – both the existence of an opinionated state of mind explaining someone's preferences can be formally demonstrated and its uniqueness up to particular class of transformations. Here however we do not want to assume completeness, since we want to be able to work with cases in which an agent forms or withdraws her preference judgements. This will pose no problem with respect to the existence of explanatory states of mind, but it will follow that they are not typically unique and that very different representations of the agent's state of mind may be consistent with what we know about her preferences. To avoid the complications that this gives rise to, we will simply assume here that opinionated state of minds are unique up to a choice of scale for measuring degrees of desire. Indeed, since the choice of zero has already been settled by the normalisation of v with respect to T, the choice of unit for v is the only remaining free parameter that we will need to worry about.

Changes of Mind Changes in an agent's preference are explained or rationalised in terms of changes to her state of mind. Consequently constraints are applied in the first instance to revisions to beliefs and desires and only derivatively to preferences. The kinematical model of revisions of states of mind presented here belongs to the class of what might be called 'perturbation-propagation' models of change; models in which the change in an agent's state of mind is viewed as a two-step process.

Stage 1 The agent changes an attitude to a particular prospect or, more generally, some set of prospects.

Stage 2 She adjusts her attitudes to all other possibilities in order to restore consistency.

The processes inducing the initial change are not themselves modelled: they might include sensory experience, deliberation, reception of a message from a reliable information source, or even hypnosis. Consequently, the normative reach of the theory is confined to saying how the agent should revise her attitudes, if the impact

of experience or deliberation is correctly represented by the stage 1 constraint. The theory says nothing about what constraints an agent *should* adopt at stage 1. That task, I take it, belongs to epistemology and moral theory.

Formally a kinematical model of revision maps prior opinionated states of mind, $\langle p, v\rangle$, to posterior ones, $\langle p^*, v^*\rangle$, as a function of the changes induced by stage 1. From this may be inferred a mapping from prior (non-opinionated) states to the posterior one, the latter just being the set of all values of the mappings from the opinionated states making up the former. The stage 1 changes are identified, not by their source, but somewhat more abstractly by the constraints that they place on the posterior attitudes. These might in principle take any number of different forms, but the sort of constraints that will be studied here include:

1. New information that X: $p^*(X) = 1$
2. New probabilities across a partition $\{X_i\}$: $p^*(X_i) = \alpha_i < 1$
3. New conditional probabilities across a partition: $p^*(X_i|A) = \beta_i$
4. New desirabilities across a partition: $v^*(X_i) = \gamma_i$
5. New conditional desirabilities across a partition: $v^*(X_i|A) = \beta_i$

The first of these constraints corresponds to the case most frequently studied in theories of belief revision, namely when the epistemic effect of interaction with the environment may be summarised by the information received by the agent. The second and third cases provide progressively more permissive representations of the effect of this interaction, allowing for cases when the agent reconsiders her probabilities for some set of prospects, or even just her conditional probabilities for them, without necessarily being able to identify the information received.

In the fourth and fifth cases interaction leads to a 'taste' change, expressed as a constraint on her posterior desirabilities, or even just her conditional desirabilities. Some comment is required here, since desirability values are scale relative and hence there is a risk of ambiguity in the claim that someone's new desirabilities take a certain value. We disambiguate the constraint $v^*(X_i) = \gamma_i$ in the following way. Assume that there exists a $\Gamma \subset \Omega$, which is such that the prior desirability of any prospect is equal to that of some element of Γ.[6] Then to say that $v^*(X) = \gamma$ is to say that the agent's new evaluation of X makes her indifferent between X and the element of Γ having prior desirability of γ (relative to the scaling of the prior desirability measure v).

It should be emphasised that these constraints on posterior probabilities and desirabilities already reflect a *judgement* on the part of the agent, some distillation of their experience or interpretation of it. These judgements may be of varying kinds and complexity and will often be completely unconscious, e.g. the interpretation of a visual stimulus as an observation of a cat. Sometimes it is possible to identify the input to the cognitive system that induces that adoption of the constraints in question, but this is by no means always the case.

[6] For instance, Γ could consists of all prospects of the form $P_i A \vee \neg P_i B$, where A and B are the most and least preferred prospects and the P_i are 'ethically neutral' prospects of varying probability.

11.3 Generalised Conditioning

We start with the most general case that we will consider, namely when the initial perturbation is exhaustively described by a redistribution of probability and desirability across a particular partition $\{A_i\}$ of the space of prospects. For instance, such a set of constraints might represent the outcome of a tasting of a range of wines or of a debate that one has had with someone on some issue. In this case we say that the agent revises her state of mind by generalised conditioning on the partition $\{A_i\}$ just in case her new state of mind $\langle p^*, v^* \rangle$ is related to prior state $\langle p, v \rangle$ by, for all prospects $X \in \Omega$:

$$p^*(X) = \sum_i p(X|A_i).p^*(A_i) \tag{11.3}$$

$$v^*(X) = \sum_i [v(X|A_i) + v^*(A_i)].p^*(A_i|X) \tag{11.4}$$

Note the similarity of these expressions to the Equations (11.1) and (11.2).

It is relatively straightforward to establish that p^* and v^* are respectively a probability and desirability function so that $\langle p^*, v^* \rangle$ is indeed a rational state of mind.[7] The interesting question, of course, is whether, or under what conditions, $\langle p^*, v^* \rangle$ is the uniquely rational state of mind to arrive at after revision when the effect of experience is correctly described by the redistribution of probability and desirability over the partition in question. The answer is straightforward: generalised conditioning is demonstrably the uniquely rational way of revising one's state of mind whenever interaction with the environment leaves one's conditional degrees of desire, given the A_i, undisturbed, i.e. whenever:

Condition 5 *(Rigidity)* $v^*(\cdot|A_i) = v(\cdot|A_i)$

Theorem 6 *(Bradley [7]) If the pairs* $\langle p, v \rangle$ *and* $\langle p^*, v^* \rangle$ *satisfy the Rigidity condition with respect to elements of the partition* $\{A_i\}$*, then* $\langle p^*, v^* \rangle$ *is obtained from* $\langle p, v \rangle$ *by generalised conditioning on* $\{A_i\}$*.*

Arguably Rigidity should hold whenever the redistribution of probability and desirability over the A_i describes all and everything that is learnt by the agent as a result of interaction with the environment and all changes to the agent's partial attitudes are rational effects of what she learns. To make this thought more precise we show that an agent whose conditional desires do not satisfy the Rigidity condition under the postulated circumstances is vulnerable to a money pump.

Consider firstly the limiting case in which experience teaches the agent that A. For the purposes of the exercise let us suppose that the truth of prospects can be bought and sold in some market so that, in an appropriate currency, $v(X)$ and $v^*(X)$ give the fair prices for the agent, before and after learning that A, of the prospect

[7] Proof in Bradley [5].

of X. Suppose firstly that the agent commits herself to a revision policy in case of learning that A such that for some X, $v(X|A) \neq v^*(X|A)$. There are two cases:

(i) $v(X|A) > v^*(X|A)$. In this case the agent can be sold the option of XA for $v(XA)$ and the option of A can be bought from her for $v(A)$. Once the option of A has been exercised the option of XA can be bought from the agent for $v^*(XA)$. By assumption $v(XA) - v(A) = v(X|A) > v^*(X|A) = v^*(XA)$. So in this case she is $v^*(X|A) - v(X|A) > 0$ poorer.

(ii) $v^*(X|A) > v(X|A)$. In this case the option of XA can be bought from the agent for $v(XA)$ and the option of A sold for $v(A)$. Once the option of A has been exercised the option of AX can sold back to the agent for $v^*(XA)$. By assumption $v^*(XA) = v^*(X|A) > v(X|A) = v(XA) - v(A)$. So in this case she is $v^*(XA) - v(XA) + v(A) > 0$ poorer.

It follows that whenever the agent commits to a revision policy for when she learns the truth of A that fails to satisfy the Rigidity condition she will find herself open to a sure loss.

To extend this argument to the more general case of a revision policy for any redistribution of probability over a partition $\{A_i\}$ consider a two-stage revision process. At the first stage, interaction with the environment induces the agent to adopt new probabilities for the elements of the partition $\{A_i\}$, without the probability of any one of them going to one. In the second stage the agent learns which of the A_i is the truth. Suppose that this process leads to a transformation of her state of mind from $\langle p, v\rangle$ to $\langle p^*, v^*\rangle$ and then to $\langle p^{**}, v^{**}\rangle$. By our previous argument for Rigidity in the context of learning the truth of a particular prospect, $v^{**}(\cdot|A_i) = v^*(\cdot|A_i)$ and $v^{**}(\cdot|A_i) = v(\cdot|A_i)$, since both the revisions from $\langle p^*, v^*\rangle$ to $\langle p^{**}, v^{**}\rangle$ and that from $\langle p, v\rangle$ to $\langle p^{**}, v^{**}\rangle$ fall under its scope. It follows that $v^*(\cdot|A_i) = v(\cdot|A_i)$ for any of the A_i and hence that the Rigidity condition holds for pure probabilistic shifts as well.

Money pump arguments, like their close relatives the Dutch Book arguments, show that failure to satisfy some condition or other renders the agent vulnerable to exploitation. It does not follow without further argument that rigidity of conditional attitude is a requirement of rationality under the given circumstances. After all, one can render oneself invulnerable to money pumps by simply not declaring a revision policy. Indeed this would seem to be a sensible precaution since there are cases in which one's attitudes may change as a result of interaction with the environment but not (entirely) because of the information that one acquires during it simply because the manner in which something is learnt has some non-rational effect on one's attitudes. If, for instance, one learns of the consequences of excessive alcohol consumption by doing the drinking oneself or of the presence of a poisonous snake in the house by standing on it, there is every possibility that other attitudes will be altered in the process and in a manner not representable as conditioning on what has been learnt. Unless the manner in which information is acquired can be controlled somehow (as perhaps it is in scientific experiments), it would be unwise to commit oneself to a revision policy in the manner required by the money-pump argument. The money pump argument is therefore inconclusive.

There are moreover reasons for thinking that no conclusive argument could be given for Rigidity, since it cannot be ruled out that the initial perturbation of the agent's attitudes provides inferential grounds for revision of conditional desirabilities. The agent might reason, for instance, that if A is less desirable than previously thought, then since AB is no less desirable than before, it must be the case that B is more desirable conditional on A than previously thought. This would lead to a violation of Rigidity.

On the other hand, should an agent's conditional desires change, either directly as a result of experience or by inference for it, the net effect on the agent's attitudes can be expressed by a set of constraints on some more refined partition. For instance, if she reasons as above, so that not only her degrees of desire for A change but also her conditional degrees of desire for B given A, then she could adopt as her posterior constraint a redistribution of probability and desirability over the partition $\{AB, A\neg B, \neg A\}$ rather than the initial partition $\{A, \neg A\}$. Then we can ask whether Rigidity is satisfied relative to this more refined partition. Crucially there will always be some level of refinement at which Rigidity will be satisfied. More exactly:

Theorem 7 *(Bradley [8]) Assume that* Ω *is countable. Let* $\langle p, v\rangle$ *and* $\langle p^*, v^*\rangle$ *be respectively an agent's prior and posterior states of mind. Then there exists some partition of* Ω *such that* $\langle p^*, v^*\rangle$ *is obtained from* $\langle p, v\rangle$ *by generalised conditioning on this partition.*

Theorem 7 tells us that any revision of a state of mind is representable as an instance of generalised conditioning as long as the revision produces a new state of mind that is internally consistent. It does not follow of course that agents, or even just the rational ones, always revise by this method or indeed that they should. But we can conclude they always revise 'as if' by generalised conditioning on a particular partition. Similarly, from the agent's own point of view, if they can adequately express the import of their experience by a redistribution of probability and desirability across some partition, then rationality requires that they adjust their attitudes to all other prospects so as to achieve a new state of mind that is related to the old one by the expressions characterising generalised conditioning.

11.4 Desire-Driven Change

Generalised conditioning is demonstrably rational whenever the Rigidity condition holds. But in order to revise one's preference in this way it is necessary to start with a rather rich input, namely an exhaustive specification of the effects of experience on one's distribution of probability and desirability across a particular partition. In this section we shall attempt to go a bit further than this and consider how a rational agent should revise her attitudes when the conditioning base is less rich. The two most salient cases are when experience gives the agent immediate cause to revise her degrees of desires, but not her beliefs, and when it gives her cause to revise her

beliefs but not her desires. In these cases the basis for conditioning takes the form of a redistribution of either probability or desirability (but not both) across some partition of the space of prospects.

In this section we consider the first of these cases, in the next section the second. Intuitively, there are two kinds of effects of desire changes that are relevant to preference revision. The first kind is the effect on the desirability of some prospect of a change in the desirability of its possible consequences. For instance, if the agent's taste in music 'matures' over time, so that her high regard for rock music and low regard for classical music is replaced by a low regard for the former and a great appreciation of the latter, then her earlier preference for a night in a club over an evening at a concert hall is likely to be reversed in time. More generally, if prospect A has greater conditional probability given X than an alternative prospect B then a rise/fall in the desirability of A relative to B should, *ceteris paribus*, lead to a rise/fall in the desirability of X.

The second kind is the effect of a change in the conditional desirabilities of a set of alternative prospects, given the presence of some condition, on the desirability of the condition itself. For instance, discovering that strawberries taste even better if eaten with cream, may lead one to value cream more highly and to purchase it more often. More generally, a rise/fall in the conditional desirability, given B, of some epistemically possible prospect A, should, *ceteris paribus*, lead to a rise/fall in the desirability of B.

To derive these intuitive conclusions concerning the effect on an agent's preferences of desire or 'taste' changes from the model of generalised conditioning, we need to make an additional hypothesis concerning the effect of desire change on belief.

Condition 8 *Global Independence of belief from desire: A change in an agent's desires should have no effect on her beliefs.*

The Global Independence principle, though seemingly natural, is no mere consistency condition. It's a condition of rational belief formation that rules out, amongst other things, wishful thinking and its pessimistic opposite, when your desiring something to be true makes it seem more or less likely to be so. Note also that it is a principle of causal (and not probabilistic) independence. A change in an agent's desires may well be evidence of a change in her beliefs, but it should never be the rational cause of them.

Even so, the condition is too strong as it stands. If the desirability of some prospect increases, for instance, I might well infer that I will try and secure its realisation. This will make the probability of my taking certain actions greater, for if I am rational then my actions will be guided by the expected desirability of their consequences. Similarly, if I have normative beliefs – beliefs about what is desirable – then these beliefs may depend on what I desire and how strongly.

There are two ways one might deal with this problem. One is to qualify the independence condition so as to exclude cases like these; this will be straightforward if exceptions to the Global Independence condition are restricted to specific categories, like the agent's beliefs about her own actions or her normative beliefs. A second is to

break revision into two steps: a first stage in which Global Independence is applied to revision in response to taste change and second stage in which beliefs are revised in response to the changes either in the expected desirabilities of actions induced by the first stage revisions (using, for instance, the kind of rules described by Skyrms [17]) or the desirabilities of prospects which are the content of the agent's normative beliefs. This second option is the one that I would advocate, but a proper exploration of it is beyond the scope of this paper.

11.4.1 Taste Change

Suppose that the agent's tastes for the elements of a partition $\{A_i\}$ change as a result of some learning experience; for instance, one of those illustrated by Cases 4, 5 and 6 in the initial list of causes of preference change. If Global Independence applies then the agent's posterior probabilities should equal her prior ones. On the other hand, her posterior desirability for any prospect X will depend on the extent to which that prospect is probabilistically connected to each of the A_i. Given the background assumption of Rigidity, the precise extent of this effect can be derived as follows from Equation (11.4) and the definition of conditional desirability:

$$\begin{aligned} v^*(X) &= \sum_i [v(XA_i) + v^*(A_i)].p(A_i|X) \\ &= \sum_i [v(X|A_i).p(A_i|X)] + [v^*(A_i) - v(A_i)].p(A_i|X) \\ &= v(X) + \sum_i [v^*(A_i) - v(A_i)].p(A_i|X) \end{aligned}$$

The term $[v^*(A_i) - v(A_i)].p(A_i|X)$ may be regarded as the measure of the desirability gain transmitted to X by virtue of the change in taste for prospect A_i and the probabilistic dependence of A_i on X. The agent's new preferences will differ from her old, in virtue of such a change of taste, as a function of the magnitude of these gains and losses. In particular her new preferences over prospects can be derived from her old plus the magnitudes of desirability transmitted, in virtue of the fact that $X \succeq^* Y$:

$$\begin{aligned} &\Leftrightarrow v^*(X) \geq v^*(Y) \\ &\Leftrightarrow v(X) + \sum_i [v^*(A_i) - v(A_i)].p(A_i|X) \geq v(Y) + \sum_i [v^*(A_i) - v(A_i)].p(A_i|Y) \\ &\Leftrightarrow v(X) - v(Y) \geq \sum_i [v^*(A_i) - v(A_i)].[p(A_i|Y) - p(A_i|X)] \end{aligned}$$

It follows that a preference reversal between X and Y will occur just in case the difference in the magnitudes of the desirability gains to each as a result of the taste change is greater than the prior difference in their desirability. For instance, to pick

up our previous example, if the A_i refer to listening to different types of music and X and Y to going to classical concert and a rock club respectively, then an earlier preference for the latter over the former will be reversed just in case the sum of the desirability gain associated with listening to classical music and desirability loss associated with listening to rock exceeds the initial desirability difference between the evening at the concert hall and the night at the club.

11.4.2 Conditional Desires

Sometimes our conditional desires change in response to experimentation or verbal testimony e.g. when we discover that particular wines are better when twinned with some types of food than others or when somebody commends a visit to a particular tourist attraction in the event that you are ever in the neighbourhood. These changes in conditional desires can occur without an initial change in the desirability of the condition itself, to the food types or to the visiting of the neighbourhood. But typically they will be inferentially relevant to them: that the food enhances the wine may be grounds to consume it and that a tourist attraction exists may be a reason to visit the neighbourhood. On the kinematical approach taken here we model these inferential effects in terms of the two stages: first there is a perturbation to the conditional desires, then these are propagated to the unconditional ones.

Formally, suppose that the agent's conditional desires, given some prospect A, for the elements of a partition $\{B_i\}$ change as a result of experience, without her desirabilities across the partition $\{A, \neg A\}$ changing. Then the constraint on her posterior state of mind is given by her new posterior conditional desirabilities $v^*(B_i|A)$ and the requirements that $v^*(A) = v(A)$ and $v^*(\neg A) = v(\neg A)$. If this represents the total initial perturbation of her state of mind, then this change may be represented by a redistribution of desirability across the partition $\{AB_i, \neg A\}$. Assuming Rigidity with respect to this partition, it follows from Equation (11.4) that:

$$v^*(XA) = v(XA) + \sum_i [v^*(B_i|A) - v(B_i|A)].p(B_i|XA)$$
$$v^*(X\neg A) = v(X\neg A)$$

(This is proved in the appendix as Lemma 11.7.1.) The term $[v^*(B_i|A) - v(B_i|A)].p(AB_i|X)$ may be regarded as the measure of the desirability gain transmitted to X by virtue of the change in the conditional desirability of B_i given A.

Consider, in particular the effect of this change on the agent's preference for and against A itself. From the above:

$$v^*(A) = v(A) + \sum_i [v^*(B_i|A) - v(B_i|A)].p(B_i|A)$$
$$v^*(\neg A) = v(\neg A)$$

Thus $A \succeq^* \neg A$:

$$\Leftrightarrow v(A) + \sum_i [v^*(B_i|A) - v(B_i|A)].p(B_i|A) \geq v(\neg A)$$

$$\Leftrightarrow \sum_i [v^*(B_i|A) - v(B_i|A)].p(B_i|A) \geq v(\neg A) - v(A)$$

It follows that a preference reversal between A and $\neg A$ will occur just in case the magnitudes of the desirability gains to A as a result of the changes in the conditional desirabilities of the B_i exceeds the prior difference in the desirability of A and $\neg A$. For example an initial disinclination for buying cream may be reversed by the discovery that strawberries taste better when smothered in it.

11.5 Belief-Driven Change

Intuitively there are two kinds of effects of belief change that are especially relevant to preferences, respectively illustrated by Cases 1 and 2 in the initial list of types of preference change. The first kind is the effect on what might be termed the instrumental value of some prospect of a change in the conditional probability, given its realisation, of prospects that matter to the agent. Thus if I learn that drinking red wine, but not white, reduces the chances of a heart attack, I may as a result come to prefer drinking red wine to white. And more generally we would expect that for any prospect B such that AB is preferred to $A\neg B$, a rise in the probability of B given A will result in a rise in the desirability of A (and vice versa).

The second kind of belief change relevant to preference is when a change in the probability of some possibility A makes the prospect of some possibility B more attractive, not because of any probabilistic dependence between the two, but because of the desirabilistic dependence of B on A. Thus if I have planned to take my children to the park if I can get away from work early enough, then learning that no rain is forecast for later in the day will make the prospect of getting off work early more attractive. This is not because the forecast affects the likelihood of getting off work, but because I prefer not to go to the park in the rain. More generally we would expect that, for any prospect B probabilistically independent of A and such that AB is preferred to $A\neg B$, a rise in the probability of A will result in a rise in the desirability of B (and vice versa).

To analyse these effects of belief change on preference we take as our point of departure the various Bayesian conditioning models for different kinds of evidence. I will not discuss the status of theories as I have done so extensively elsewhere,[8] save to point out that, no less than the preference kinematical models, their validity depends on the correct identification of the initial 'perturbation' of the agent's beliefs induced by interaction with the environment. There are three types of belief revision

[8] In [6].

that we want to consider: classical conditioning on information, Jeffrey conditioning on probabilistic information and Adams conditioning on conditional information.

11.5.1 Conditioning on New Information

Suppose that as a result of some such interaction with the environment an agent learns that A, and nothing more than that A, so that the initial effect of this interaction is exhaustively described by the constraint on her posterior state of mind that $p^*(A) = 1$. In these circumstances classical Bayesianism requires that the agent's new degrees of belief, p^*, should go by her initial conditional degrees of belief given that A, i.e. for all $X \in \Omega$:

$$p^*(X) = p(X|A)$$

How should the agent revise her desires in this case? Notice first that it follows from the axiom of averaging that:

$$v^*(X) = v^*(XA).p^*(A|X) + v^*(X\neg A).p^*(\neg A|X) = v^*(XA)$$

In particular, $v^*(T) = v^*(A)$, and so by the axiom of normality, $v^*(A) = 0$. Hence $v^*(X) = v^*(XA) - v^*(A) = v^*(X|A)$. It follows that if the Rigidity condition is satisfied, so that the agent's posterior desirabilities equal her prior conditional desirabilities given that A, then for all $X \in \Omega$:

$$v^*(X) = v(X|A)$$

Whenever an opinionated agent revises her beliefs and desires in this manner – by conditioning on the information that A – her new preferences between any prospects X and Y will go by her old preferences between the prospects XA and YA, i.e.:

$$X \succeq^* Y \Leftrightarrow XA \succeq YA$$

which is what we would expect whenever learning that A is the case has a purely informational effect on the agent's preferences. For example, in Case 2 of our initial list of causes of preference change, my prior preference for the prospect of my neighbour holding a barbecue in the rain to that of his holding it in sunny weather, is what explains my acquisition of a preference for rain over sunny weather upon receipt of the information that he has planned a barbecue.

11.5.2 Probabilistic Updating

Conditioning on acquired information is not the only kind of belief revision of relevance to preference change. Sometimes our probabilities for prospects change even

though there is no definite proposition that we newly regard as certain and which can serve an adequate basis for classical Bayesian conditioning. For instance one may glimpse someone in the distance, but not be certain that it is the friend one is expecting, or think that one remembers being told something, without being certain about it. In these cases, and others, the effect of experience is best represented by a redistribution of probability across a set of mutually exclusive and exhaustive prospects rather than by the stock of newly acquired information.

Suppose that $\{A_i\}$ is just such a partition and that as a result of interaction with the environment (or indeed reflection or deliberation), the agent's probabilities for each A_i changes from $p(A_i)$ to $p^*(A_i)$. In these circumstances, an agent obtains her new degrees of belief p^*, by Jeffrey conditioning on the partition $\{A_i\}$ just in case p^* is related to p by the 'kinematical' formula:

$$p^*(X) = \sum_i p(X|A_i).p^*(A_i)$$

i.e. by computing her new probabilities for any prospect X by averaging her old conditional probabilities for X using her new unconditional probabilities for the A_i.

How should the agent revise her degrees of desire in these circumstances? As in the case of desire-driven change, a further hypothesis about the effect of belief changes on desires is required in order to derive the answer from the generalised conditioning model. In general, changes in belief concerning some prospects should and will affect the desirability of other prospects. But a change in the degree to which one believes that X does not in itself give one reason to change one's attitude to the desirability of X itself. And more generally changes in the relative probability of the elements of some partition of prospects do not rationalise changes in their relative desirability. For instance, suppose that I strongly prefer to teach an advanced course in decision theory than an introductory course in logic. Then getting wind of the Head of Department's intentions regarding teaching allocations should not make any difference to the degree to which my preference for the former alternative exceeds the latter. To be sure, changes in belief regarding some prospect can cause a change its desirability without being a reason for it. This is what happens in cases of adaptive preference change (like that illustrated by Case 3 in our list of preference change types) when, for instance, the fact that some outcome seems less likely causes the agent to desire it less (or more). Such adaptations are rule out by the following condition.

Condition 9 *Local Independence of preference from belief: A change in an agent's degrees of belief over a partition $\{A_i\}$ should have no effect on her relative preferences for the A_i, i.e.:*

$$A_i \succeq^* A_j \Leftrightarrow A_i \succeq A_j$$

The Local Independence condition will not suffice to determine a unique expression for the $v^*(A_i)$ in terms of p, v and the $p^*(A_i)$ – given the scale-dependence of the desirability measures this is too much to expect – but it does significantly constrain it. I would contend that the simplest and most natural expression satisfying

the principle and the axioms of desirability and normality is the following:

$$v^*(A_i) = v(A_i) - k \tag{11.5}$$

where $k = \sum_i v(A_i).p^*(A_i)$ expresses the desirability gain to the agent as result of the change in probabilities (informally, we can say that it expresses the amount by which the world has been proved, by the experience inducing the belief-change, to be a better or worse place than initially believed).

If both the Local Independence condition, as expressed in Equation (11.5), and the Rigidity condition hold, then the agent must compute her new desirabilities for any prospect X by averaging her old desirabilities for the elements of the partition $\{XA_i\}$ of X by her new conditional probabilities for the A_i given X, and then renormalising. For under these assumptions it follows from the generalised conditioning model that:

$$v^*(X) = \Sigma_i v(XA_i).p^*(A_i|X) - k \tag{11.6}$$

where as above $k = \sum_i v(A_i).p^*(A_i)$. Renormalisation is required to ensure that the agent's posterior desirabilities satisfy the axiom of normality, but has no significance for her posterior preferences.

When an opinionated agent revises in the described manner then her new preferences will be related to her old by:

$$X \succeq^* Y \Leftrightarrow \Sigma_i v(XA_i).p^*(A_i|X) \geq \Sigma_i v(YA_i).p^*(A_i|Y)$$

i.e. her new preferences over prospects will go by her new expectations of (old) desirability, given their realisation.

Example: My friend sometimes serves lunch al fresco, depending on whether he can be bothered to set the table outside but (rather idiosyncratically) independently of what the weather is like. Since I expect poor weather and I don't like eating outside in these conditions I am hoping that he will not be bothered. The weather report turns out to be much more positive however and so I revise my probabilities for warm weather on the day. Let B be 'eat outside' and A be 'warm weather'. Since by assumption A is probabilistically independent of B:

$$B \succeq^* \neg B \Leftrightarrow v(BA).p^*(A) + v(B\neg A).p^*(\neg A) \geq v(\neg BA).p^*(A) + v(\neg B\neg A).p^*(\neg A)$$

I will reverse my preference for $\neg B$ over B just in case the revision of belief transfers sufficient probability, conditional on B, to the desirable prospects. This change is illustrated in the table below, showing my prior preferences (when $p(A)$ is low) and posterior ones (when $p^*(A)$ is high). More preferred prospects appear above those that are less preferred.

Table 11.1 Change in belief

$< p, v >$: $p(A)$ is low	$< p^*, v^* >$: $p^*(A)$ is high
AB	AB
$\neg A \neg B$	$\neg A \neg B$
$\neg B$	B
T	T
B	$\neg B$
$A \neg B$	$A \neg B$
$\neg AB$	$\neg AB$

11.5.3 Change in Conditional Belief

Not just our partial beliefs, but our conditional beliefs too can change as a result of a number of different kinds of interaction with our environment, including observation, experimentation and testimony. For example, we may receive verbal testimony in the form of a conditional statement or data about relative frequencies in a population. Such changes are not always naturally represented as consequences of learning the truth of some set of evidence propositions and so cannot readily be assimilated to the classical conditioning model presented above.

In a particularly interesting set of cases of this kind interaction with the environment gives us cause to change one or more of our conditional beliefs, given some possibility, without it giving us cause to change our probabilities for the possibility itself. Formally, let $\{B_i\}$ be a partition of propositions such that $1 > p(B_i|A) > 0$ and suppose that the agent is caused to change her conditional degrees of belief for the B_i given A from $p(B_i|A)$ to $p^*(B_i|A)$. Then her new partial beliefs, p^*, are said to be obtained from p by Adams conditioning on this change in conditional probabilities just in case:

$$p^*(X) = p(X \neg A) + \sum_i p(XAB_i).\frac{p^*(B_i|A)}{p(B_i|A)}$$

What makes Adams conditioning particularly salient is the fact that, in a certain sense, it is the exact complement of Jeffrey conditioning. For in Adams conditioning it is the conditional probabilities with respect to elements of a partition that change while the probabilities of the elements themselves remain rigid, rather than the other way round. Consequently study of this kind of revision offers the possibility of extending kinematical modelling to cases where interaction with the environment affects both the agent's unconditional beliefs and her conditional ones, by representing them in terms of combinations of Jeffrey and Adams conditioning.

How should an agent revise her preferences in this case? We can gain purchase in this question by recognising that Adams conditioning on this change in conditional belief is formally identical to Jeffrey conditioning on the more refined partition $\{AB_i, \neg A\}$ under the additional constraint that the agent's degree of belief in A

is unchanged.[9] Given this, and the assumption that the Rigidity condition applies to the more refined partition, it will follow by application of Equation (11.6) to this partition, that the agent's new degrees of desire for any prospect X should be obtained from her old by averaging her old degrees of desire for the XAB_i by her new conditional degrees of belief for the AB_i given X (and renormalising). More formally:

$$v^*(X) = \sum_i v(XAB_i).p^*(AB_i|X) + v(X\neg A).p^*(\neg A|X) - k$$

where $k = \sum_i v(AB_i).p^*(B_i|A).p(A) + v(\neg A).p(\neg A)$.

We are often interested in the relevance of the change in the conditional probabilities of the B_i given A to the desirability of A itself; for instance when A is an action that might be performed and the B_i are the possible consequences of its performance. It follows from the above that:

$$v^*(A) = \sum_i v(AB_i).p^*(B_i|A) - k$$
$$v^*(\neg A) = v(\neg A) - k$$

Hence:

$$A \succeq^* \neg A \Leftrightarrow \sum_i v(AB_i).p^*(B_i|A) \geq v(\neg A)$$

Example: I am considering the prospect (A) of an invitation to lunch at a friend. From past experience I know that if we are invited he will serve either (B) take-away pizza, which I rather like, or (C) homemade Lasagne, which I can barely stomach. I am unsure as to whether to accept or not, but then I am reminded by my wife that our friend served Lasagne last time we went for lunch and this makes it almost certain that pizza will be served. As a result, the prospect of an invitation appears a good deal more attractive. Schematically we have the following reversal of preference between A and $\neg A$ as a result on my change in conditional belief.

Table 11.2 Change in Conditional belief

$< p, v >$: $p(B\|A)$ is low	$< p^*, v^* >$: $p^*(B\|A)$ is high
AB	AB
$\neg A$	A
T	T
A	$\neg A$
AC	AC

[9] For a proof see Bradley [6. Theorem 1].

11.6 Preference Loss and Preference Gain

In this final section I want to look at cases in which interaction with the environment leads an agent to acquire a new preference or to withdraw a preference judgement, rather than revise it. To do so we must now drop the working assumption of the previous section that agent's prior and posterior preferences are complete. Instead we suppose that her state of mind is non-opinionated and hence represented by a (non-singleton) set of pairs of probability and desirability functions, $S = \{\langle p_i, v_i\rangle\}$. I shall refer to each of these pairs as an avatar of the agent, so the we can say such things as that one avatar desires one prospect more than another, while another avatar does not. Whatever constraints interaction with the environment place on the agent's posterior state of mind must now be reflected in revisions to the states of mind of all her avatars.

A preference for one prospect A over another B is gained by an agent (in this framework) whenever (1) her prior state of mind is such that it is neither the case for all her avatars i, $v_i(A) \geq v_i(B)$ nor that for all such i, $v_i(B) \geq v_i(A)$, and (2) interaction with the environment induces a change in the agent's state of mind such that her for all her avatars i, $v_i^*(A) \geq v_i^*(B)$. A preference for prospect A over another B is lost just in case the opposite is true, i.e. initially all avatars value A more than B, but after revisions some (but not all) value B more highly. It remains for me to show that this can happen. I will do so by means of examples.

Preference Gain Suppose that the agent has no prior preference between prospects X and Y, but that for some good prospect G, it is the case that $XG \approx YG > X\neg G \approx Y\neg G$. Suppose furthermore that observation induces a revision to her beliefs such that for all avatars i, $p_i^*(G|X) = p_i^*(\neg G|Y) = 1$. If each avatar i revises by Adams conditioning then its posterior desirabilities will be such that:

$$\begin{aligned} v_i^*(X) &= v_i^*(XG).p_i^*(G|X) + v_i^*(X\neg G).p_i^*(\neg G|X) \\ &= v_i^*(XG) \end{aligned}$$

and:

$$\begin{aligned} v_i^*(Y) &= v_i^*(YG).p_i^*(G|Y) + v_i^*(Y\neg G).p_i^*(\neg G|Y) \\ &= v_i^*(Y\neg G) \end{aligned}$$

But $v_i^*(XG) > v_i^*(Y\neg G)$. Hence $v_i^*(X) > v_i^*(Y)$.

Preference Loss Suppose that the agent has just two avatars, 1 and 2, such that $p_1(G) = 0.9$, $p_2(G) = 0$, $v_1(GX) > v_1(GY) > v_1(\neg GY) > v_1(\neg GX)$ and $v_2(\neg GX) > v_2(\neg GY)$. Then $X > Y$ since both avatars value X more highly than Y, albeit for different reasons. Now suppose that the agent observes that $\neg G$. Then if the two avatars revise their attitudes by conditioning, the second's state of mind will remain the unchanged while $p_1(G) = 0$, $v_1^*(X) = v_1(\neg GX) - v_1(\neg G) < v_1(\neg GY) - v_1(\neg G) = v_1^*(Y)$. So it will no longer be the case that both avatars value X more highly than Y, despite the fact that they share more information. Hence the preference for X over Y is lost.

11.7 Appendix

Lemma 11.7.1. *Let $\langle p, v\rangle$ and $\langle p, v^*\rangle$ be two pairs of probability and desirability functions such that:*
(i) Rigidity over $\{AB_i, \neg A\}$: $v^*(X|AB_i) = v(X|AB_i), v^*(X|\neg A) = v(X|\neg A)$
(ii) Independence: $v^*(A) = v(A), v^*(\neg A) = v(\neg A)$
Then:

$$v^*(XA) = v(XA) + \sum_i [v^*(B_i|A) - v(B_i|A)].p(B_i|XA)$$
$$v^*(X\neg A) = v(X\neg A)$$

Proof. By the axiom of desirability, $v^*(XA) = \sum_i v^*(XAB_i).p(AB_i|X)$. Hence, by application of the definition of conditional desirability:

$$\begin{aligned} v^*(XA) &= \sum_i [v^*(X|AB_i) + v^*(AB_i)].p(B_i|XA) \\ &= \sum_i [v^*(X|AB_i) + v^*(B_i|A) + v^*(A)].p(B_i|XA) \\ &= \sum_i [v(X|AB_i) + v^*(B_i|A) + v(A)].p(B_i|XA) \end{aligned}$$

by application of the Rigidity and Independence conditions. Hence, once again applying the definition of conditional desirability:

$$\begin{aligned} v^*(XA) &= \sum_i [v(XAB_i) - v(AB_i) + v^*(B_i|A) + v(A)].p(B_i|XA) \\ &= \sum_i [v(XAB_i) + v^*(B_i|A) - v(B_i|A)].p(B_i|XA) \\ &= \sum_i v(XAB_i).p(B_i|XA) + \sum_i [v^*(B_i|A) - v(B_i|A)].p(B_i|XA) \\ &= v(XA) + \sum_i [v^*(B_i|A) - v(B_i|A)].p(B_i|XA) \end{aligned}$$

by the axiom of desirability. On the other hand, by application of the definition of conditional desirability:

$$\begin{aligned} v^*(X\neg A) &= v^*(X|\neg A) + v^*(\neg A) \\ &= v(X|\neg A) + v(\neg A) \\ &= v(X\emptyset A) \end{aligned}$$

by application of the Rigidity and Independence conditions.

References

1. Becker, Gary. 1996. *Accounting for Tastes*. Cambridge, MA: Harvard University Press.
2. van Benthem, Johan and Fenrong Liu. 2007. The Dynamics of Preference Upgrade. *Journal of Applied Non-Classical Logics* 17: 157–182
3. Bolker, Ethan. 1966. Functions Resembling Quotients of Measures. *Transactions of the American Mathematical Society* 124: 292–312
4. Bowles, Samuel. 1998. Endogenous Preferences: The Cultural Consequences of Markets and Other Economics Institutions. *Journal of Economic Literature* 36: 75–111
5. Bradley, Richard. 1999. Conditional Desirability. *Theory and Decision* 47: 23–55
6. Bradley, Richard. 2005. Radical Probabilism and Mental Kinematics. *Philosophy of Science* 72: 342–364
7. Bradley, Richard. 2007. The Kinematics of Belief and Desire. *Synthese* 156: 513–535
8. Bradley, Richard. 2009. Becker's Thesis and Three Models of Preference Change. *Politics, Philosophy and Economics* 8(2): 223–242.
9. Grüne-Yanoff, Till. 2007. Why Don't You Want to Be Rich? Preference Explanations on the Basis of Causal Structure. In *Topics in Contemporary Philosophy: Explanation and Causation*, eds. J.K. Campbell and M. O'Rourke, Cambridge, MA. London: MIT Press.
10. Hansson, Sven-Ove. 1995. Changes in Preferences. *Theory and Decision* 38: 1–28
11. Hansson, Sven-Ove. 2001. *The Structure of Values and Norms*. Cambridge: Cambridge University Press
12. Jeffrey, Richard C. 1983. *The Logic of Decision*, 2nd ed, Chicago IL: University of Chicago Press
13. Jeffrey, Richard C. 1992. *Probability and the Art of Judgement*. Cambridge: Cambridge University Press
14. Liu, Fenrong. 2008 *Changing for the Better: Preference Dynamics and Agent Diversity*. Ph.D. thesis, Institute for Logic, Language and Computation
15. Pettit, Philip. 2002. Decision Theory and Folk Psychology. In *Rules, Reasons, and Norms: Selected Essays*, ed. Philip Pettit, 192-221. Oxford: Oxford University Press
16. Pollak, Robert A. 1976. Interdependent Preferences. *The American Economic Review* 66: 309–320
17. Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation*. Cambridge MA: Harvard University Press
18. Stigler, G. J. and Gary S. Becker. 1977. De Gustibus Non Est Disputandum. *The American Economic Review* 67: 76–90

Chapter 12
Population-Dependent Costs of Detecting Trustworthiness: An Indirect Evolutionary Analysis*

Werner Güth, Hartmut Kliemt, and Stefan Napel

Abstract If the (un)trustworthy are rare, people will talk about them, making their detection more reliable and/or less costly. When, however, both types appear in large numbers, detecting (un)trustworthiness will become considerably more difficult and possibly too costly to provide a positive feedback supporting preferences underlying trustworthy behavior. We analyze how the composition of a population of trustworthy, respectively, untrustworthy individuals evolves if the cost and reliability of type detection depend on the population composition.

12.1 Introduction

If virtuous behavior prevails, a rare misdeed will be conspicuous. It will become a matter of gossip and widely known. If nearly everybody is misbehaving, the rare trustworthy individual will tend to raise a lot of interest, too. Again behavior may become widely known. In short, bad as well as good conduct may stand out in a crowd of behavior of the other kind. It will easily be observed and thereby trigger responses of observers (see Coleman (1988) on such mechanisms from a social science point of view). These responses in turn may feed back on the process in which the preferences underlying the behavior itself are adapted to the "habitat" of the actors.

W. Güth (✉)
Max Planck Institute of Economics, Kahlaische Strasse 10, 07745 Jena/Germany
e-mail: gueth@econ.mpg.de

H. Kliemt
Frankfurt School of Finance and Management
e-mail: hartmut.kliemt@t-online.de

S. Napel
Department of Economics, University of Bayreuth
e-mail: stefan.napel@uni-bayreuth.de

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

To study the population dependency of detecting virtue we focus on the virtue of being trustworthy, respectively of failing to show this moral quality. By showing trust the trustor aims at reaching a payoff dominant result as compared to the status quo of no trust but makes himself vulnerable to an act of "exploitation" by the trustee. Trustworthiness is modeled as a modification of the preferences of the trustee. As a result of this modification (due to some kind of intrinsic motivation or preference) the trustee evaluates results in ways other than suggested by objective or material outcomes that reflect "reproductive" success in the context of the evolutionary model. Intrinsic "moral preferences" prevent the trustworthy trustee from exploiting the trustor, whereas untrustworthy individuals will not refrain from exploitation should they be trusted.

We assume that to limit their risk, trustors can invest in type detection. Utilizing such a technology they receive a type signal whose reliability and cost are, however, not constant as in Güth and Kliemt (2000) but which can depend on the population composition. We initially investigate what to expect when given reliabilities require higher costs of detection for more symmetrically composed populations.[1] An extension of our analysis allows that the reliability of the signal of another's type may depend on how the population is composed. More specifically, we assume that the signal reliabilities (a signal's reliability when resulting from the trustworthy may differ from that stemming from the untrustworthy) become worse when the relative frequencies of both types – as characterized by their different preferences for virtuous or non-virtuous behavior – converge.

On a more abstract level, our analysis is comparable to evolutionary studies that assume that the rules of the game change when (average) population play changes. So, for instance, Joosten et al. (2003) are studying games whose payoff parameters depend also on past play. In principle, we do the same. But in our approach the reliability and cost of detection evolve with the population composition and are in this sense endogenous to our model. Section 12.2 describes the basic setup more formally. The rational decision behavior for all possible compositions of the population with (un)trustworthy individuals is derived in Section 12.3. Assuming success-monotonic evolutionary dynamics we determine in Section 12.4 the evolutionarily stable population compositions and their basins of attraction. Section 12.5 explores several extensions before Section 12.6 puts things into perspective.

[1] One could also have assumed that not only the cost of investing in type detection is population dependent (in the sense of being lower the more one type prevails) but that also the strength of preference modifications depends on how the population is composed. If, for instance, feelings of guilt increase when one is the rare untrustworthy, this should stabilize universal trustworthiness. Similarly, if feelings of guilt get weaker when untrustworthiness becomes more common, a monomorphic society of potential exploiters should emerge.

12.2 The Model

To capture the trust problem in social interaction, we rely on the trust game in Fig. 12.1. To make our results comparable with earlier work we use the same parameter normalization with $1 > r > s > 0$ as in Güth and Kliemt (2000).

The interpretations of the moves are

N – no trust (in player 2)
T – trust (in player 2)
E – exploiting (player 1's trust)
R – rewarding (player 1's trust)

The payoffs in Fig. 12.1 (top player 1, bottom player 2) are "objective" in that they represent material or reproductive success. It is assumed, however, that individuals evaluate results (plays of the game) not only in "objective" terms. In the second–mover role their behavioral choices can be guided by preferences other than furthering reproductive success. It is the presence or absence of such preferences of sufficient strength that renders an individual trustworthy or untrustworthy, respectively. This is captured by a purely subjective payoff component m (see Fig. 12.2). There is no objective payoff corresponding to this. If $m = \underline{m} < r - 1$, we are dealing with a trustworthy type. Correspondingly, we assume for the untrustworthy type of player 2 that $m = \overline{m} > r - 1$. One interpretation of m is that of an intrinsic inclination of reciprocity in the sense of responding in kind, here by R(ewarding) T(rust), which, of course, could be implied by an internalized norm.

Let us briefly comment on this distinction between material and non-material payoffs which is a possibility but no necessity for our indirect evolutionary approach. As the indirect evolutionary approach allows to combine rational deliberation and evolutionary adaptation, one has in principle to distinguish between

- Decision utility, i.e., for all deliberated choices one has to anticipate their utility effects and select the utility best behavior, and

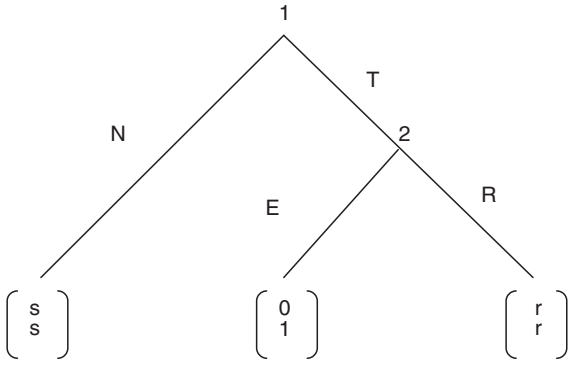

Fig. 12.1 A trust game

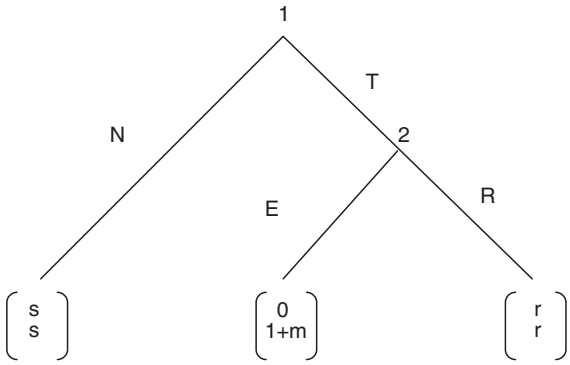

Fig. 12.2 A trust game with a subjective payoff component

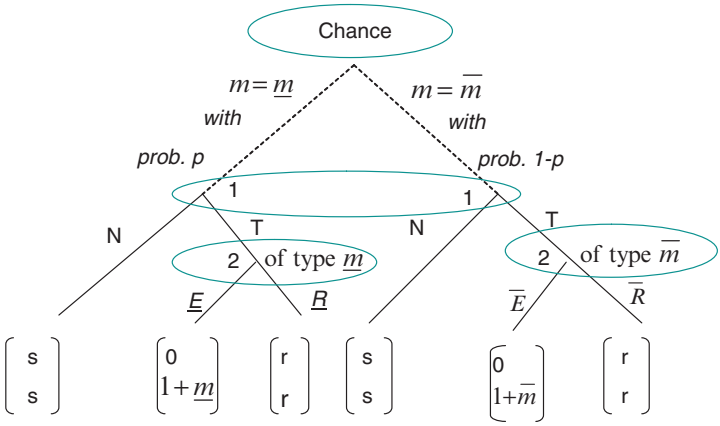

Fig. 12.3 A Bayesian trust game

- Evolutionary success which measures fitness of the evolving types, e.g. reproductive success in evolutionary biology and whatever determines the future frequencies of types in (cultural) evolution

For the example ahead it is assumed that m as well as the material payoffs define decision utility but that evolutionary success is only determined by the material payoffs, i.e., the latter differs from the former by m. Nevertheless, m influences (rational choice) behavior and thereby indirectly evolutionary success.

The actual play of the trust game is embedded in a more complex decision process as indicated in Fig. 12.3.

In Fig. 12.3 information sets are indicated by encircling all decision nodes between which the deciding player (1) cannot distinguish, i.e., where she confronts identical choices (N or T) whereas player 2 can decide in a type-dependent way, namely by choosing $\underline{R}$ when $m = \underline{m}$ and $\overline{E}$ in case of $m = \overline{m}$. Assuming an infinite population with random matching (for an alternative, see Güth et al. 2002) the selected pair of individuals i and j ($\neq i$) confronts the following basic scenario:

- Let p with $0 \leq p \leq 1$ denote the population share of individuals with preference parameters m of type $\underline{m}$ whereas the population share of untrustworthy types $m = \overline{m}$ is $1 - p$.
- Although preference types are private information, i.e., only the individual herself is aware whether her preference parameter m satisfies $m = \underline{m}$ or $m = \overline{m}$, the population shares p and $1 - p$ of trustworthy, respectively untrustworthy types are commonly known and thus determine the corresponding beliefs about the other's type as captured by the Bayesian game in Fig. 12.3.
- In the tradition of the indirect evolutionary approach, which allows to combine rational deliberation of certain choices and evolutionary adaptation or justification of other behavioral aspects (see Berninghaus et al. 2003, who compare various such combinations for the example at hand), it is assumed that the choices in the Bayesian game of Fig. 12.3 are rationally deliberated but that the population composition, as captured by the population share p of trustworthy preference types $m = \underline{m}$, evolves, i.e., increases when trustworthy types are materially more successful than untrustworthy ones and vice versa.
- Players can rationally decide to avoid playing the Bayesian game in Fig. 12.3 by investments in type detection whose costs C are population dependent in the sense of $C = C(p)$ and whose reliability may be more or less perfect.

More specifically, every pair of matched individuals encounters the following decision sequence:

- Being aware of p but not of the other's m-type, the two individuals independently decide between investing (y), resp. not investing (n) in type detection; where the cost
$$C\,(p) = \frac{1}{2}\left[kp\,(1-p) - \frac{1}{6}k + c\right] \text{ with } 6c \geq k > 0, \qquad (12.1)$$
of choosing y is population dependent. Clearly, such a cost function has the property that rare types are more cheaply found out due to $p\,(1-p) \to 0$ both for $p \to 0$ as well as for $p \to 1$. Parameter k scales the sensitivity of the detection cost to the population composition (higher k corresponds to greater population dependence). Parameter c serves as an overall cost measure because averaging $C(p)$ over all possible population states yields $\int_0^1 C\,(p)\,dp = \frac{c}{2}$ independently of sensitivity parameter k. Note that the term $c - \frac{1}{6}k$ approaches 0 for $k \uparrow 6c$ and equals the positive constant c for $k = 0$; it is thus guaranteed that $C(p) \geq 0$ for all $p \in [0, 1]$ and $6c \geq k > 0$.
- Chance assigns roles independently of player type, i.e. either individual i becomes player 1 and j player 2 or vice versa, each with probability $1/2$.
- Player 1 decides between N (no trust, which would end the interaction with what may be seen as the status quo payoffs) or T(trust, which may be seen as an invitation to co-operate). If player 1 has chosen y before, he can base his decision on a type signal $\hat{m}$ of player 2's true m-type. The reliability of that signal is determined by two parameters

$$\text{Prob}\,(\hat{m} = \overline{m} | m = \overline{m}) = \overline{\mu} \in (1/2, 1] \text{ and Prob}\,(\hat{m} = \underline{m} | m = \underline{m}) = \underline{\mu} \in (1/2, 1]$$

meaning that a truly untrustworthy $m = \overline{m}$ − type is revealed by $\hat{m} = \overline{m}$ with probability $\overline{\mu}$ larger than 1/2 whereas the misleading signal $\hat{m} = \underline{m}$ results with the complementary probability. Similarly, for a true type $m = \underline{m}$ the signal $\hat{m} = \underline{m}$ is more likely than the misleading one. Asymmetric reliabilities in the sense of $\underline{\mu} \neq \overline{\mu}$ allow to explore which risk is more decisive, the one of not trusting a trustworthy partner or the one of trusting somebody untrustworthy (see Güth and Kliemt 2000).

- In case of 1's decision for T, player 2 finally chooses between E and R.

The order of the first two decision stages is without any restriction. Reversing the order of moves and letting players decide between y and n only when actually being in the role 1 of trustor would merely divide detection costs $C(p)$ by 2, without any other changes since the trustor would only choose between investing (y) and not investing (n) when actually playing the role of player 1 and having to trust someone whose m-type she does not know (see Fig. 12.3). As the probability of becoming player 1 is $1/2$, the reversed order would result in the same level of (expected) costs if we dropped the factor $1/2$ in the equation defining $C(p)$. In the setting envisioned here it may seem more natural, though, to assume that detection costs are borne as a kind of sunk cost beforehand. People either bear the costs of following up what is going on in the group or not. When they by chance encounter a potential partner they must decide "on the spot" whether or not to engage him in a co-operative venture by showing trust or not.

The payoffs are the ones in Fig. 12.2 minus the costs $C(p)$ of type detection for the individual(s) having chosen y. These (phenotypical) payoffs determine the optimal decision behavior in the process above which will be derived in Section 12.3. Compared to this the composition of types, i.e. the evolution of the population composition parameter $p \in [0, 1]$, is governed by the objective success of the $\overline{m}$, resp. $\underline{m}-$ types. Choice behavior is governed by the subjective utility function and the parameter m in it. The objective contribution of overt behavior to evolutionary success can be read off from the payoff function by setting $m = 0$. We will analyze the evolutionarily stable population compositions p in Section 12.4.

12.3 Rational Play as Depending on the Population Composition

Player 2's behavior will depend on his type whenever he is asked to move, i.e. after the move T by player 1. More specifically, an $\overline{m}$ -type would choose E and an $\underline{m}-$ type R. Regarding player 1, it has been assumed that the population share p of trustworthy $\underline{m}-$ types is commonly known and thus the common prior for meeting a trustworthy player 2 when not investing (n) and for Bayesian updating when investing (y) in costly type detection.

After n, i.e. when not having invested in type detection, player 1 chooses T (yielding pr) rather than N (yielding s for sure) if $p \geq s/r$, and N otherwise.

After y, i.e. when having received a signal $\hat{m}$ about player 2's m-type, player 1 will follow the recommendation of a signal $\hat{m} = \underline{m}$ and choose T provided that

$$p \geq \frac{(1-\overline{\mu})\, s}{\underline{\mu}\,(r-s) + (1-\overline{\mu})\, s} : = RHS \tag{12.2}$$

where the right-hand side (RHS) of the inequality is smaller than s/r due to $\overline{\mu}$, $\underline{\mu} > 1/2$ and $s < r$.

If p is below this threshold level, even a signal $\hat{m} = \underline{m}$ indicating the trustworthiness of player 2 cannot convince player 1 given her pessimistic initial beliefs about the chances that trust will be rewarded. Similarly, for a sufficiently optimistic prior (large p), even a 'bad' signal $\hat{m} = \overline{m}$ cannot dissuade player 1 from trusting. Given her updated beliefs after $\hat{m} = \overline{m}$ she chooses N only provided that

$$LHS : = \frac{\overline{\mu} s}{\overline{\mu} s + \left(1-\underline{\mu}\right)(r-s)} \geq p. \tag{12.3}$$

where the left-hand side (LHS) above is larger than s/r due to $\overline{\mu}$, $\underline{\mu} > 1/2$ and $s < r$.

Costly detection activity can be profitable only if the signal is not discarded, i.e. if its recommendation is followed always – and not only when it matches the intended action based on the prior. Thus further analysis of investment will focus on intermediate values of p satisfying both of the above conditions (see Güth and Kliemt 2000, Lemma 3.1). Such values exist since the RHS and LHS are smaller and larger than s/r, respectively.

Now consider the initial choice between y and n. Optimal behavior *after* n yields the payoff expectation

$$pr \quad \text{for} \quad p \geq s/r \tag{12.4}$$

$$s \quad \text{for} \quad p < s/r \tag{12.5}$$

conditional on being assigned to the role of player 1.

Choosing y and afterwards always following the recommendation yields

$$\frac{1}{2} p \underline{\mu} r + \frac{1}{2}\left[p\left(1-\underline{\mu}\right) + \overline{\mu}\,(1-p)\right] s - C\,(p) \tag{12.6}$$

plus a constant term capturing payoff in case the considered agent is allocated to the role of player 2 (which cannot be influenced by the agent's n or y-decision). Investigating when Equation (12.6) exceeds $pr/2$ for $p > s/r$, and $s/2$ respectively for $p < s/r$, it follows that y is better than n (or at least as good) for the subinterval

$$\overline{p}\,(C\,(p)) \geq p \geq \underline{p}\,(C\,(p)) \tag{12.7}$$

of $LHS \geq p \geq RHS$ with

$$\overline{p}\,(C\,(p)) = \frac{\overline{\mu} s - 2C\,(p)}{\overline{\mu} s + \left(1-\underline{\mu}\right)(r-s)} \tag{12.8}$$

(derived from case $pr \geq s$) and

$$\underline{p}\,(C\,(p)) = \frac{(1-\overline{\mu})\,s + 2C\,(p)}{\underline{\mu}\,(r-s) + (1-\overline{\mu})\,s} \tag{12.9}$$

(derived from case $pr < s$). It is possible that $\overline{p}\,(C\,(p)) < \underline{p}\,(C\,(p))$, and then investing in type detection (the decision y) is necessarily suboptimal. This case arises whenever

$$2C\,(p) > \left(\underline{\mu} + \overline{\mu} - 1\right)(r-s)\,\frac{s}{r}. \tag{12.10}$$

However, if average cost $c/2$ and population sensitivity k are not too large, $\overline{p}\,(C\,(p)) \geq p \geq \underline{p}\,(C\,(p))$ will be satisfied for an entire interval of population compositions.[2] This is illustrated in Fig. 12.4. The triangle depicts $\overline{p}\,(\kappa)$ and $\underline{p}\,(\kappa)$ for different detection cost levels κ (in the range $\overline{p}\,(\kappa) \geq \underline{p}\,(\kappa)$, given $r = 0.8$, $s = 0.4$, $\underline{\mu} = \overline{\mu} = 0.85$), and different curves $\kappa = C\,(p)$ illustrate how increasing population sensitivity ($k \in \{c, 6c\}$ with $c = 0.1$) affects the range of p such that indeed $\overline{p}\,(C\,(p)) \geq p \geq \underline{p}\,(C\,(p))$. In our view, both situations, i.e., those where k is too large and those allowing for a generic interval, where type detection pays, are relevant. If the stakes of the trust exchange are rather low, it usually will not pay to invest in finding out the m-type of one's partner. Compared to that, high-stake exchanges will usually render costly type detection a reasonable investment when the population is not too monomorphic. It seems advantageous that our approach allows for both situations.

12.4 The Evolution of the Population Composition

Whenever p violates $\overline{p}\,(C\,(p)) \geq p \geq \underline{p}\,(C\,(p))$, i.e. when not investing in type detection (n) is optimal, T will be chosen by player 1 if $p \geq s/r$ and N otherwise. For $p \geq s/r$ an $\underline{m}$ – type will receive material reward r whereas an $\overline{m}$ – type is in objective payoff terms more successful. Hence for $p \geq s/r$, any monotonic evolutionary dynamics imply that p decreases as long as $p \geq s/r$ (and T is optimal).

Suppose that in the range where investment in type detection does not pay, p at some point starts to satisfy $p < s/r$. Then player 1 chooses N, and both m-types fare equally in the second mover role, too. But even then, if there are "trembles" in the sense of rare unintentional choices by player 1, the decline of p will continue (see Selten 1983,1988). This goes on until either $p^* = 0$ is reached or until p arrives in the range where $\overline{p}\,(C\,(p)) \geq p \geq \underline{p}\,(C\,(p))$ and y becomes optimal.

[2] A necessary and sufficient condition for this is that $\overline{p}\,(C\,(p)) > \underline{p}\,(C\,(p))$ holds at $p = s/r$, i.e. at the peak of the $\underline{p}\,(\kappa) - \overline{p}\,(\kappa)$-triangle in Fig. 12.4. This amounts to $k\frac{s}{r}\left(\frac{r-s}{r}\right) - \frac{k}{6} + c < \left(\underline{\mu} + \overline{\mu} - 1\right)(r-s)\,\frac{s}{r}$, where the left-hand side can be made arbitrarily small through an appropriate choice of c and k.

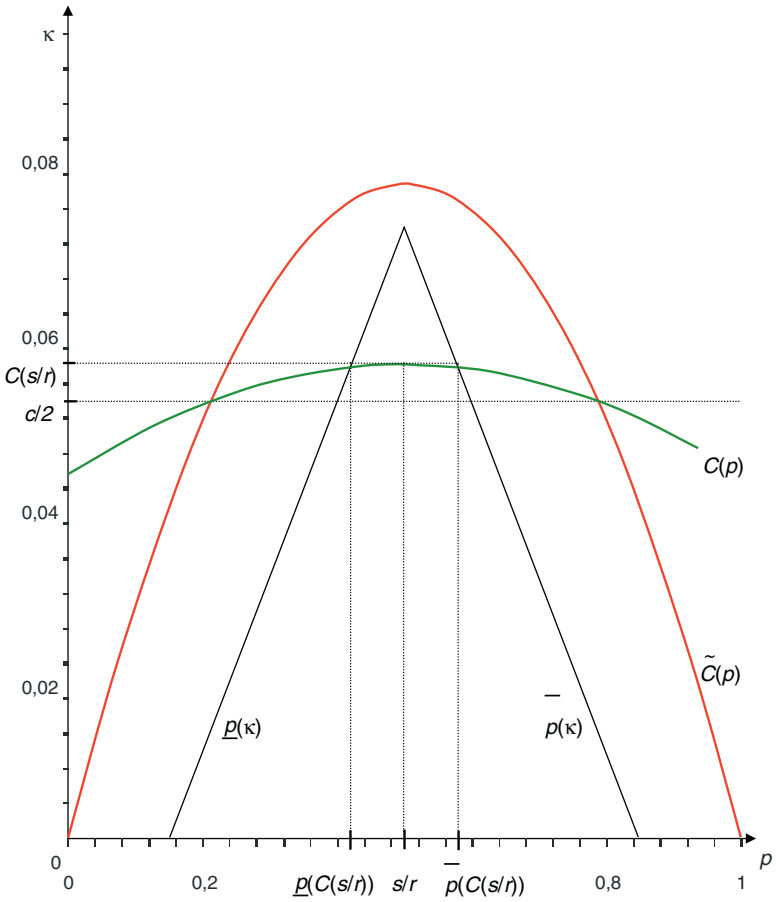

Fig. 12.4 Costs and benefits of type detection

So, consider population compositions p satisfying $\overline{p}\,(C\,(p)) \geq p \geq \underline{p}\,(C\,(p))$, i.e. player 1 invests in type detection and follows the signal $\hat{m}$ which he receives. Here the material payoff of player 2 depends on his m-type as follows:

$$\underline{\mu} r + \left(1 - \underline{\mu}\right) s - C\,(p) \;\text{ for } m = \underline{m} \tag{12.11}$$

$$1 - \overline{\mu} + \overline{\mu} s - C\,(p) \;\text{ for } m = \overline{m}. \tag{12.12}$$

The trustworthy $\underline{m}-$ type fares better than the unreliable $\overline{m}-$ type of player 2 if

$$\frac{\underline{\mu}}{1 - \overline{\mu}} > \frac{1 - s}{r - s} \tag{12.13}$$

and vice versa if the opposite inequality applies (the degenerate case of equality is negligible). Both possibilities emerge in generic parameter regions. The positive difference $r - s$ is what each player gains by playing (T, R) rather than N whereas the larger difference $1 - s$ is what the exploiting type would gain by being trusted. For given material payoffs of the game, the right-hand side of Equation (12.13) is thus a constant larger than 1. Thus, when signals become rather unreliable in the sense of $\underline{\mu}, \overline{\mu} \searrow 1/2$, the inequality in Equation (12.13) will typically be violated whereas for more and more perfect signals $\left(\underline{\mu}, \overline{\mu} \nearrow 1\right)$, it will usually hold.

The reverse of Equation (12.13) is true when $\overline{m}-$ types are likely to be mistaken for a trustworthy $\underline{m}$ – type (low $\overline{\mu}$) and can then realize a substantial gain. Then $\overline{m}$ – types fare universally better and p will sooner or later – faster when $\overline{p}\,(C\,(p)) \geq p \geq \underline{p}\,(C\,(p))$ or $p > s/r$, and slower otherwise – decrease to $p^* = 0$ which is for these parameter configurations the only evolutionarily stable population composition.

Whenever the inequality in Equation (12.13) holds (typically for $\overline{\mu}$ close to 1) p increases in the range $\overline{p}\,(C\,(p)) \geq p \geq \underline{p}\,(C\,(p))$, which depends on p due to the population dependency of $C\,(p)$, and decreases outside this range. This, of course, means that for all initial population compositions $p_0 < \underline{p}\,(C\,(p_0))$ or $p_0 > \overline{p}\,(C\,(p_0))$ one starts out with a decrease of p over time whereas for $\overline{p}\,(C\,(p_0)) \geq p_0 \geq \underline{p}\,(C\,(p_0))$ one starts with an increase of p over time. In the latter case this process will finally lead to a stable population composition p satisfying

$$\overline{p}\,(C\,(p)) = p \tag{12.14}$$

or

$$kp^2 - \left[\overline{\mu}s + \left(1 - \underline{\mu}\right)(r - s) + k\right] p + \overline{\mu}s - c + \frac{k}{6} = 0. \tag{12.15}$$

Note that $\overline{p}(C(p))$ decreases (increases) for $p < 1/2$ $(>1/2)$, and $\overline{p}(C(0)) = \overline{p}(C(1)) < 1$. Therefore, of the two solutions of the quadratic equation

$$p = \frac{\overline{\mu}s + \left(1 - \underline{\mu}\right)(r - s) + k}{2k} \pm \frac{\sqrt{\left[\overline{\mu}s + \left(1 - \underline{\mu}\right)(r - s) + k\right]^2 - 4k\left(\overline{\mu}s - c + \frac{k}{6}\right)}}{2k}, \tag{12.16}$$

only the smaller one qualifies as an evolutionarily stable population composition.

How does one actually determine the direction in which p changes (up or down) when considering a time point $t = 0$ at which the population composition is $p_0 \in [0, 1]$? Let us describe this for the more interesting case in which Equation (12.13) applies and p would increase in the interval $(\underline{p}\,(C\,(p_0)), \overline{p}\,(C\,(p_0)))$. The first thing to check is whether $2C\,(p_0) > (\underline{\mu} + \overline{\mu} - 1)(r - s)\frac{s}{r}$ holds at all. If not, the interval $(\underline{p}\,(C\,(p_0)), \overline{p}\,(C\,(p_0)))$ is empty and p would decrease (fast or slow) throughout. If the condition holds, however, we know that $p_0 \in (\underline{p}\,(C\,(p_0)), \overline{p}\,(C\,(p_0)))$ and that p will increase from p_0 to a new level p_t, for which one repeats the analysis.

So, if $\overline{p}\left(C\left(p_0\right)\right) \geq p_0 \geq p\left(C\left(p_0\right)\right)$, the population composition converges from below to

$$\overline{p}^* = \frac{\overline{\mu}s + \left(1-\underline{\mu}\right)(r-s) + k - \sqrt{\left[\overline{\mu}s + \left(1-\underline{\mu}\right)(r-s) + k\right]^2 - 4k\left(\overline{\mu}s - c + \frac{k}{6}\right)}}{2k}. \tag{12.17}$$

If instead

$$1 \geq p_0 > \overline{p}\left(C\left(p_0\right)\right), \tag{12.18}$$

then p converges to $\overline{p}^*$ from above because for $p > \overline{p}\left(C\left(p_0\right)\right)$ nobody invests in type detection (i.e. chooses n). Now $p_0 > \overline{p}\left(C\left(p_0\right)\right)$ is equivalent to

$$kp_0^2 - \left[\overline{\mu}s + \left(1-\underline{\mu}\right)(r-s) + k\right]p_0 + \overline{\mu}s - c + \frac{k}{6} < 0 \tag{12.19}$$

or $p_0 > \overline{p}^*$. In view of this, the basin of attraction for $\overline{p}^*$ is $p_0 > \underline{p}\left(C\left(p_0\right)\right)$, whereas the basin of attraction for $p^* = 0$ is $p_0 < \underline{p}\left(C\left(p_0\right)\right)$.

Condition $p_0 > \underline{p}\left(C\left(p_0\right)\right)$ (noting that $\underline{p}\left(C\left(p\right)\right)$ first increases and then decreases in p and that $\underline{p}\left(C\left(0\right)\right) = \underline{p}\left(C\left(1\right)\right) > 0$) can also be expressed as

$$p_0 > D := \frac{k - \underline{\mu}(r-s) - (1-\overline{\mu})s + \sqrt{\left[k - \underline{\mu}(r-s) - (1-\overline{\mu})s\right]^2 + 4k\left[s(1-\overline{\mu}) + c - \frac{k}{6}\right]}}{2k}. \tag{12.20}$$

Accordingly there exists a threshold D determining whether an initial population composition p_0 leads to an $\overline{m}-$monomorphism or $p^* = 0$, namely for $p_0 < D$, or to a bimorphic population composed of a $\overline{p}^*$-share of $\underline{m}-$ types and a complementary $1 - \overline{p}^*-$ share of $\overline{m}-$ types, namely when $p_0 > D$.

Dynamics for given population-dependent costs $C(p)$ are illustrated in Fig. 12.5. The solid line indicates comparatively "fast" movement, corresponding to a strict payoff (dis)advantage of trustworthy agents. Movement along the dotted line is "slow" because it is driven by mutations, i.e. agents in the role of player 1 who by mistake trust and then make trustworthy agents in the role of player 2 fare worse than others (who take advantage of the mistake).

It can easily be seen that a stronger sensitivity of investment costs $C(p)$ to changes in the population composition, i.e. a higher coefficient k, increases D and thus the basin of attraction of $p^* = 0$. At the same time, it also decreases $\overline{p}^*$. Therefore, if the costs of type detection rise faster as the rarer m-type gets less rare (costs are "more" population dependent), the chances that a bimorphic population emerges are worsened and the bimorphic population will on average be characterized by a lower p or be less "virtuous".[3]

[3] The effect of c, i.e. of the fixed cost parameter, has already been discussed by Güth and Kliemt (2000); see also their discussion of the reliability parameters $\underline{\mu}$ and $\overline{\mu}$.

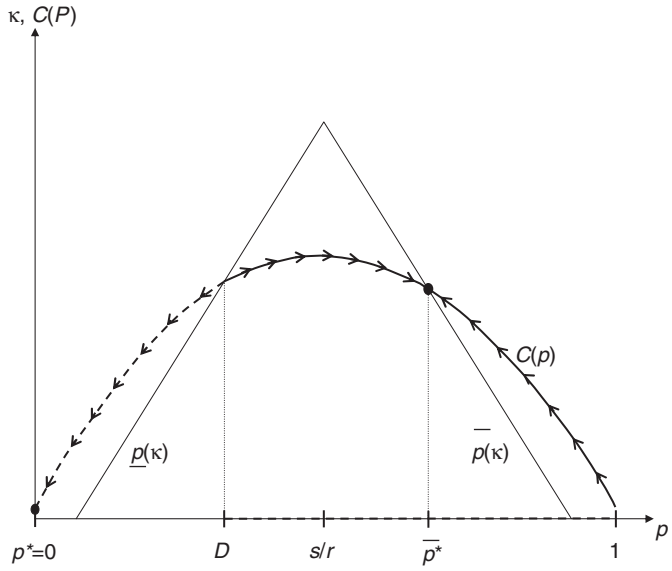

Fig. 12.5 Cost and population dynamics

12.5 Extensions

The preceding model of population-dependent detection costs lends itself to a number of variations and extensions. First, the specific, simple functional form of $C(p)$ could be substituted by another one. Many alternatives to the analytically convenient quadratic shape exist. For instance, a bell shaped form of $C(p)$ might be plausible under certain circumstances. This would reflect that costs may initially increase only slowly and then more sharply as more and more (un)trustworthy individuals are added to a population dominated by $\underline{m}-$ types ($\overline{m}-$ types) and eventually reverse. This case allows for multiple stable polymorphic population states, as illustrated in Fig. 12.6. Of the six intersection points of $C\,(p)$ and the triangle shown in the figure only $\overline{p}_1^*$, $\overline{p}_2^*$ and $\overline{p}_3^*$ are stable bimorphisms with generic basins of attraction (indicated by the direction arrows on $C\,(p)$), whereas $\underline{p}_1$, $\underline{p}_2$ and $\underline{p}_3$ are merely watersheds separating different basins of attraction.

Second, as already indicated in the introduction, population-dependent costs of detection can be viewed as capturing in an indirect way the population dependency of *signal reliabilities*. So an alternative setup of the model would have taken detection costs to be fixed at some level c but assumed that reliability parameters $\underline{\mu}$ and $\overline{\mu}$ decrease (in a possibly asymmetric fashion) as the population state p approaches its least informative level of $1/2$. Different plausible versions of population-dependent reliabilities $\underline{\mu}\,(p)$ and $\overline{\mu}\,(p)$ could then be considered. While many of them would merely induce a rounded version of the $\overline{p}\,(\kappa)$ and $\underline{p}\,(\kappa)$-triangle in Figs. 12.4– 12.6 (with evolution of p along a horizontal line $C(p) = \kappa$), new phenomena may also arise. In particular, it is possible that there are more than two preference reversals

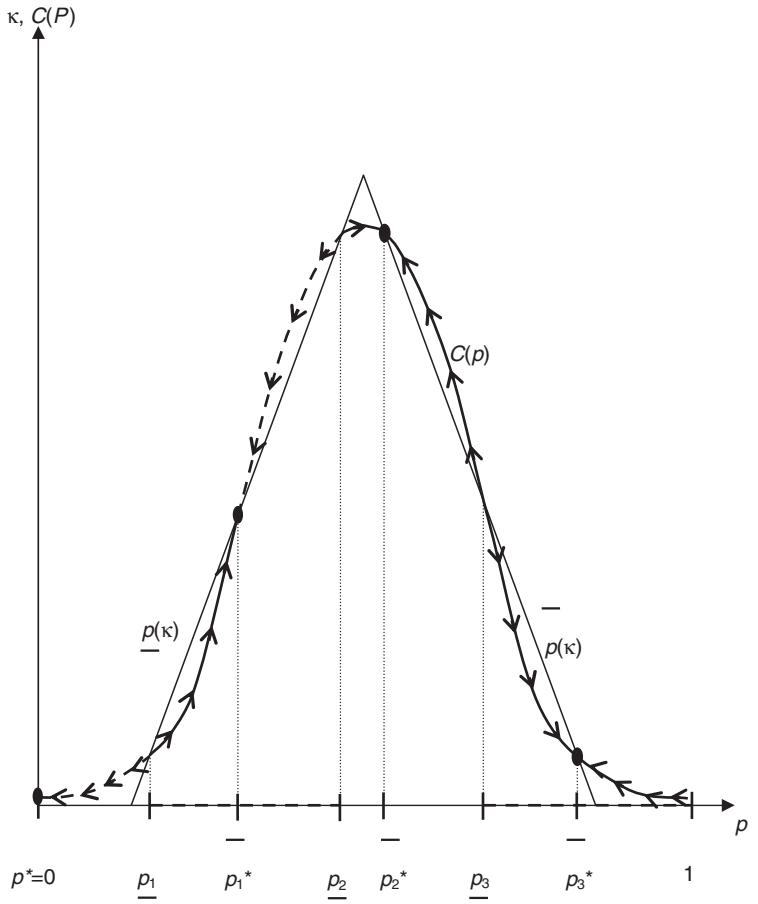

Fig. 12.6 Dynamics for bell shaped detection costs

regarding the decision of whether to invest in type detection (y) or not (n) as the share of trustworthy agents, p, increases from 0 to 1. This is illustrated in Fig. 12.7 (using parameters $r = 0.7$, $s = 0.35$, $k = 1.8$, and $\mu^{\max} = 0.99$) for the case of

$$\underline{\mu}\,(p) = \overline{\mu}\,(p) = \begin{cases} \max\left\{\mu^{\max} - kp^2, \frac{1}{2}\right\} & \text{if } p \leq \frac{1}{2}, \\ \max\left\{\mu^{\max} - k\,(1-p)^2, \frac{1}{2}\right\} & \text{if } p > \frac{1}{2}. \end{cases} \tag{12.21}$$

These (symmetric) reliabilities fall from a maximal level of $\mu^{\max}$ at $p = 0$ and $p = 1$ to $\mu^{\max} - k/4$ or $1/2$, whichever is larger, as $p = 1/2$ is approached. So the probability $1 - \underline{\mu}$ of falsely expecting an untrustworthy $\overline{m}$ – type in case of an $\underline{m}$ – encounter is lower (higher) when $\overline{m}$ – types are rare (close to $1/2$-population share), and the probability $1 - \overline{\mu}$ of falsely expecting an untrustworthy $\underline{m}$ – type in case of an $\overline{m}$ – encounter is smaller (larger) when $\underline{m}$ – types are rare (close to $1/2$-population share).

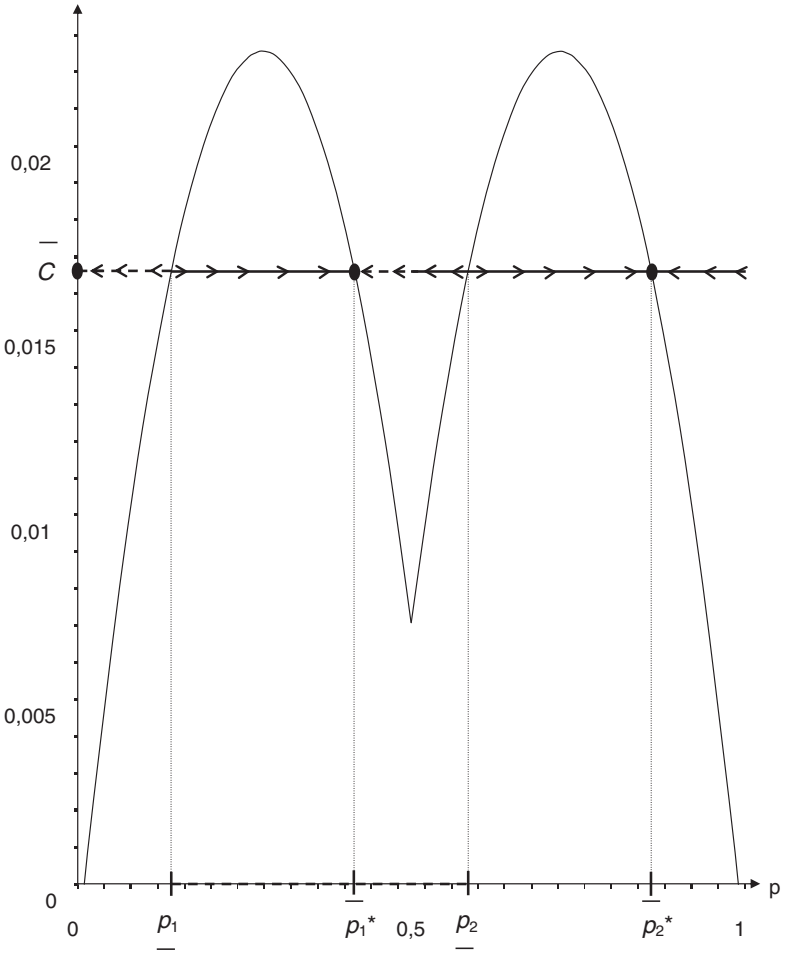

Fig. 12.7 Dynamics for population-dependent signal reliabilities

For the simple cost function of $C(p) \equiv \overline{C}$ with $\overline{C}$ small enough to lie below the two peaks (see Fig. 12.7 – note that, in general, the two peaks can have different heights) the result can be described as follows[4]: the two stable bimorphisms $\overline{p}_1^*$ and $\overline{p}_2^*$ have generic basins $(\underline{p}_1, \underline{p}_2)$ and $(\underline{p}_2 1)$, respectively; i.e., $\underline{p}_1$ and $\underline{p}_2$ are again watersheds separating the basins of attraction of $\overline{p}_1^*$, $\overline{p}_2^*$ and the stable monomorphism $p^* = 0$ with $(0, \underline{p}_1)$ as its basin of attraction.

One can obviously complicate the analysis and create more stable bimorphisms, e.g., by combining a bell-shaped cost function as in Fig. 12.6 with two-peaked

[4] We do not explicitly distinguish between fast and slow decline of p above the camel-shaped curve indicating gross benefits from exploiting type signals: the decline is driven by mutations only for $s/r < p$, while for $s/r > p$ there is a strict disadvantage for $\overline{m}$ – types even without rare mutations.

"triangles" as in Fig. 12.7. However, it seems that such specific speculation is much less fruitful than the general insight that by allowing population-dependent type detection costs and/or reliabilities of type signals,[5] the phenomenon of multiple stable bimorphisms can arise. It is thus possible that equally structured societies (e.g. societies with more or less equal payoff parameters as defining the interaction structure in Section 12.2) reveal different positive population shares of (un)trustworthy individuals, solely because they started out differently. It also suggests a new kind of policy for raising the level of trustworthiness in a society which can be described as "basin or watershed jumping". In pursuit of such a policy a measure could, for instance, aim at restarting the evolutionary p-process above the lower watershed of the preferred bimorphism and then let evolution run its course.

It is worth noting, though, that aiming at a $p = 1$-monomorphism (except by assuming $\overline{\mu} = \underline{\mu} = 1$ and $C(1) = 0$), see also Fn. 1 above) does not make sense. Moreover the emergence of a $p^* = 0$-monomorphism can never be ruled out entirely. It seems that in an imperfect world we have to live with some untrustworthy individuals at any rate and, should we ever happen to have an unfortunate start of social interaction with too few trustworthy individuals, initially, we may end up with a population of untrustworthy individuals only. If so, there is a reason for policy intervention. A policy trying to induce an evolutionary increase of p, may succeed in reaching a stable population composition with self-sustaining levels of trustworthiness. And then, depending on how the policy maker assesses the basic structure of the interaction another intervention may "jump start" the interaction to a new evolutionary path in a "better basin of attraction". Obviously there may be costs to such policy interventions. Depending on those costs and the likelihood of policy success or failure it may be wise or not to aim at reaching the bimorphism with a maximally stable population share of individuals who are endowed with a preference for showing trustworthy behavior.

12.6 Putting Things into Habitual Perspective

Whether or not so-called "ecological rationality" (see Gigerenzer and Todd 1999; Smith 2003) is rationality in the strict sense may be contested. However, it seems rather clear that average human behavior is adapted, and adapts by several alternative mechanisms, to behavioral ecological niches. Since the same individuals interact in different contexts – each forming a kind of behavioral ecological niche – with the same or other individuals (see also Aumann and Güth 2000) it is not at all obvious which kinds of (habitual) population compositions might be stable if levels of success in different contexts influence compositions.

[5] Another alternative (with similar qualitative implications as the cases already considered) would be to allow agents to optimally *choose* the desired individual signal reliabilities according to a cost function $C\left(\overline{\mu}, \underline{\mu}, p\right)$ which is increasing in $\overline{\mu}$ and $\underline{\mu}$, and decreasing in distance $\left|\frac{1}{2} - p\right|$.

Should type detection become the easier the more monomorphic the population becomes this could be interpreted inversely within the context of a discussion of "habitual ecology": Living in the habitat becomes cognitively more demanding or complex when the numbers of trustworthy and untrustworthy individuals become more equal. In this sense our rather specific analysis can be interpreted as an exemplary demonstration of how the population composition and the quality of a habitat may be co-evolving.

The possibility of multiple stable bi-morphisms should make us think twice, too, as far as policy measures are concerned. More often than not we tend to think of policy interventions in terms of permanent interference and regulation of interaction. As opposed to that, an effort to "jump start" social evolution without sustained intervention may be the better policy. Such measures as subsidizing type recognition or type signaling may be sufficient to enhance trustworthiness initially. After reaching a more favorable basin of attraction we can let the "natural" self-sustaining process run its course. Once the visible hand of government is used it may let go and hand over the improvement of behavior and the adaptation of the underlying preferences to the invisible one again. Hopes that interventionist policies might impose such self-restraint on themselves may, however, be futile.

On a more fundamental philosophical level the preceding shows how difficult the implementation of normative ethical standards in moral practices and moral institutions may be. It seems that the moral philosopher is as much at fault as the political economist when he conceives his role as that of an advisor to a benevolent despot. In both cases full control over the social process in which some normative ideal or other is to become real is a dangerous illusion. Political ideals must be implemented in social reality in ways that take into account the constraints of actual human behavioral dispositions and motivations. The same holds good for moral ideals. The moral ideals must somehow become incorporated into mental processes, emotions etc. of actual human actors and express themselves in the preferences of the individuals concerned.

In short, how moral suggestions based on ideal ethical theories work themselves out in practice depends on institutional factors and regularities in, or frequencies of human behavior. Our corresponding moral practices of ascribing responsibility, praise and blame are strongly influenced by the frequency of the behavior that is evaluated. The preceding discussion illustrates the underlying social evolutionary processes by a specific case of morally relevant behavior and demonstrates that preferences and cognitive processes supporting such behavior may depend in a rather intricate manner on its prevalence in a population.

References

Aumann, R. and W. Güth. 2000. Species survival and evolutionary stability in sustainable habitats – The concept of ecological stability. *Journal of Evolutionary Economics* 10: 437–447.

Berninghaus, S., W. Güth, and H. Kliemt. 2003. From teleology to evolution: Bridging the gap between rationality and adaptation in social explanation. *Journal of Evolutionary Economics* 13(4): 385–410.

Coleman, J. S. 1988. Free riders and zealots: The role of social networks. *Sociological Theory* 6: 52–57.

Gigerenzer, G. and Peter M. Todd (eds.). 1999. *Simple heuristics that make us smart*. New York: Oxford University Press.

Güth, W. and H. Kliemt. 2000. Evolutionary stable co-operative commitments. *Theory and Decision* 49: 197–221.

Güth, S., W. Güth, and H. Kliemt. 2002. The dynamics of trustworthiness among the few. *The Japanese Economic Review* 53(4): 369–388.

Joosten, R., T. Brenner, and U. Witt. 2003. Games with frequency-dependent stage payoffs. *International Journal of Game Theory* 31: 609–620.

Selten, R. 1983. Evolutionary stability in extensive two-person games. *Mathematical Social Science* 5: 269–363.

Selten, R. 1988. Evolutionary stability in extensive two-person games – correction and further developments. *Mathematical Social Science* 16: 223–266.

Smith, V. L. 2003. Constructivist and ecological rationality in economics. *American Economic Review* 6: 465–508.

Index