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:

$$des(p) = \frac{\sum\limits_{\{w \in W | p \in w\}} prob(w) \times des(w)}{\sum\limits_{\{w \in W | p \in w\}} prob(w)}$$

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 $A \succ_t$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 $\succeq_\mathrm{A}$ 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

→

	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.