Noname manuscript No.
(will be inserted by the editor)

Mistakes or Reflexive Preferences?

João V. Ferreira

Abstract Neoclassical economics uses the maximization of a stable and exogenous preference relation as the benchmark for positive and normative economics. Following the evidence that behavior depends on context and experience, several authors have designed models that treat behavior inconsistent with the maximization of a stable preference as mistaken. In this paper, I argue that it is important to distinguish mistakes from inconsistent behavior that results from preferences (or preference change) that individuals identify with (*reflexive preferences*). I sketch two hierarchical preferences models that represent some of these ideas, and discuss how they could relate with the conflicted agent and evolving agent models in order to represent reflexive and non-reflexive preference change. Finally, I argue that collecting information on if individuals identify or not with their preferences (or preference change) may be useful for normative analysis, in particular as a refinement to welfare rankings currently used in behavioral welfare economics.

Keywords: Reflexive preferences; Mistakes; Preference change; Hierarchical preferences; Behavioral welfare economics.

JEL classification: B4; D03; D6.

1 Introduction

At least since Arrow (1951), it has been standard practice in neoclassical economics to assume that all tastes, values, or other preferential considerations of an individual can be summarized in a single ordering over all relevant alternatives.[1] In most economic theory and application, this ordering is taken to be exogenous and stable over time. Positive economics has viewed the maximization of this single preference as the main driving force underlying individual behavior. In normative economics, these preferences are the main ingredients for evaluating the desirability of alternative states of affairs.

The findings of psychology and behavioral economics show however that models based on an exogenous and stable preference relation are often at odds with the dependence of behavior on context and experience.[2] As a response, many authors have designed models that treat choice behavior that is inconsistent with the maximization of a stable preference as errors or mistakes.[3] These models often assume that individuals have a “true” underlying preference that they would follow would their reasoning not been distorted by a faulty psychological mechanism, or, in the absence of such an assumption, that the best for these individuals would be to follow such “consistent” preference relation.

J. V. Ferreira
Aix-Marseille Univ., CNRS, EHESS, Centrale Marseille, & AMSE.
Centre de la Vieille Charité, 2, Rue de la Charité, 13002, Marseille, France.
E-mail: joao.vferreira@univ-amu.fr

[1] In Arrow (1951, 17) “[i]t is assumed that each individual in the community has a definite ordering of all conceivable social states [alternatives], in terms of their desirability to him. It is not assumed here that an individual’s attitude toward different social states is determined exclusively by the commodity bundles which accrue to his lot under each. It is simply assumed that the individual orders all social states by whatever standards he deems relevant.” I thank Nicolas Gravel for this reference.

[2] See Hoff and Stiglitz (2016) for a recent review and taxonomy of many of these findings and the strands of literature associated with them.

[3] See Rabin (2013) for a recent review.

In this paper, I argue that it is important to distinguish between the "inconsistent" behavior that results from preferences (or preference change) that the individuals *identify with*[4] and the "inconsistent" behavior that results from preferences that the individuals do not identify with. The difference is that while the later are preferences that the individuals do not endorse (and that lead to choices that are judged as mistakes by the individuals themselves), the former are preferences that the individuals endorse even if they lead to behavior that is inconsistent with the maximization of a stable preference relation. I argue that this distinction may be useful for positive economics, since it may lead to better description and prediction of economic behavior. Since preferences that individuals identify with are based on their evaluation of themselves, in what follows I call them *reflexive preferences*.

I sketch two hierarchical (preferences) models that represent some of these ideas. The hierarchical *retrospective model* takes a backward-looking perspective, where preferences (choices) are judged by the individual *ex-post*. The hierarchical *evolving model* takes a forward-looking perspective, where choices are the result of reflexive or non-reflexive preferences. "Inconsistent" behavior may result from reflexive or non-reflexive preference change, i.e., preference change that an individual identifies with or preference change that an individual does not identify with respectively. I argue that behavior that is inconsistent with the maximization of a stable preference but that individuals identify with (reflexive preference change) is *not* the result of a mistaken psychological mechanism.

I then distinguish two representations of the economic agent that could be coupled with a hierarchical model to represent reflexive and non-reflexive preferences and preference change. One is based on (i) the conflict between multiple preferences and the other on (ii) the evolution of a single preference. The first refers to the cases in which an individual identifies (or not) with several motivations, roles, or goals that lead her to "alternate" between different preferences. The second refers to the cases in which an individual changes her single preference over time according to her experience and identifies (or not) with these changes.

This view is intimately linked with a *person*'s reflexive ability to form preferences over preferences, or what philosophers often call second-order desires, volitions, or preferences.[5] I argue that data on second-order preferences may be useful for normative analysis, since it adds information on what people *identify* with, what people *value*, what they *care* about, and *who* they wish to be or become. I discuss how such information can be used as a *refinement* to welfare rankings currently used in behavioral welfare economics (e.g. Bernheim and Rangel 2007, 2009), and rejoin some of the criticisms to the use of second-order preferences in economics.

The remainder of the paper is organized as follows. I start with a review of some of the arguments in favor of taking the maximization of a stable and context-independent preference relation as a benchmark in positive economics, notably the ones put forward by Hausman (2012) in support of sticking to a notion of preference as a *total subjective comparative evaluation* (Section 2). I follow with a review of the strategy of extending this approach to include mistakes (Section 3), and argue that it is important to distinguish between preferences that individuals identify with and preferences they don't identify with (Section 4). I then formalize some of these notions in Section 5, and discuss two conceptions of the economic agent that could be used to represent reflexive and non-reflexive preferences and preference change (Section 6). I continue with an appraisal of the use of preferences over preferences as a tool for welfare analysis (Section 7), and discuss some of the features, limitations, and extensions of this approach in Section 8. I conclude with a brief comment (Section 9).

2 The Rational Agent

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

In neoclassical economics' textbooks and most theory and application agents are assumed to have stable and exogenous preferences over all relevant alternatives. This means that whether an agent prefers

[4] An individual identifies with a preference when (roughly) she evaluates or judges this preference positively, endorses it, and/or wants it to be her will. I discuss the philosophical basis of this notion in Section 8.2. To avoid awkward wording, I refer to individuals in the feminine.

[5] See Frankfurt (1971) for the philosophical basis of higher-order desires and volitions. See Jeffrey (1974) for a first treatment of second-order preferences. See e.g. Sen (1977) and George (1984) for treatments of preferences over preferences in economics.

x to y remains stable over time and across contexts, and that preferences are taken to be an essential but unexplained feature of the economic agent's identity. In particular, the agent never changes his or her "true" fundamental preferences over fully specified outcomes.[6] If $\mathcal{T} = \{1, ..., T\}$ denotes a sequence of periods and $\succsim$ denotes such preference, this means that if $x \succsim y$ in period t then $x \succsim y$ for every other $t' \in \mathcal{T}$, where the statement $x \succsim y$ can be read as "x is preferred or indifferent to y". Stigler and Becker (1977, 76) illustrate this view: according to the authors, economists should "treat tastes [preferences in their paper] as stable over time", and "search for differences [changes] in prices or incomes to explain any differences or changes in behavior".

The observational implications of this model are described by the traditional revealed preference axioms.[7] For instance, the Weak Axiom of Revealed Preference (WARP) states that if an alternative x is "revealed preferred" to y (i.e., x is once chosen when y is available *and* rejected), then y is *not* revealed to be "at least as good as" x (i.e., y is never chosen when x is available). This axiom is necessary and sufficient for a choice function to be rationalized by the maximization of a single preference (see e.g. Sen 1971). In terms of consumption decisions (i.e., choices over budget sets), the Generalized Axiom of Revealed Preferences (GARP, a stronger condition than WARP) is necessary and sufficient for "rational" behavior (see Afriat 1967).

In a recent book, Daniel Hausman (2012) provides an appraisal of the economists' rational agent model that has attracted a lot of attention in the literature (see e.g. Infante et al. 2016). He is interested in describing how the concepts of *preference, value, choice, and welfare* are and ought to be used in economics, and provides, in my view, an interesting reason-based conception of preference (even though, as it will become clear, I endorse an alternative way to conceptualize preference). The author argues that the concept of a single preference, as employed in neoclassical economics, is (and ought to be) a total subjective comparative evaluation (TSCE). It is *comparative* in the sense that people prefer one state of affairs to another. It is *subjective* in the sense that the comparison is made from a first-person perspective. And it is *total* in the sense that it is a comparison that takes into account everything that the economic agent considers to be relevant for choice. In the words of Baigent (1995, 92), who shares the same view as Hausman (2012) on this point, "[w]hat is being assumed is that agents who have multiple cares and concerns have resolved any conflicts into an 'all-things-considered preference'." It seems clear that preferences are comparative and subjective, and nowadays, I think most economists would agree that the concept of preference, as used in neoclassical economics, is most often an all-things considered ranking of alternatives.[8]

But according to Hausman (2012) a preference is (and ought to be) also an *evaluation*, in the sense that it is the result of a rational deliberation about what agents have most reason to do. Hausman (2012) argues that a preference is (or should be seen as) a reason-based evaluation rather than a judgment, rather than an expression of taste, or rather than a feeling, because judgments do not by themselves motivate action, tastes do not exhaust the considerations relevant to choice, and feelings - alone - do not provide *reasons* for action (see also Hausman 2013, 219). This means that an agent's choice that is not based on a rational deliberation about what she has most reason to do, such as a choice based on *intuition* (defined as an automatic impression), is not considered to reveal a preference according to Hausman's (2012) definition. As I will try to motivate in what follows, it is possible to keep the advantages of seeing a preference as a total subjective comparative evaluation even if the evaluation is not necessarily based on a rational deliberation.

According to Hausman (2012, 14-20) taking preference to be a TSCE is consistent with two implicit assumptions imposed on preferences in economics (and that are relevant to my analysis).[9] First, Hausman (2012) argues that the fact that preferences take all relevant considerations for choice into account accords with the long-held idea in economics that preferences determine and motivate choices (when, as Hausman stresses, combined with beliefs). The assumption that preferences determine choices means that among

[6] The neoclassical approach is compatible with changes in preferences over *uncertain prospects* following an update in the agent's beliefs about the likelihood of the possible outcomes of those prospects. To be self-contained, I mostly abstract from questions related to risk and uncertainty.

[7] Examples of these axioms include the Weak and Strong Axioms of Revealed Preference and the Weak and Strong Congruence Axioms. See Sen (1971) for a seminal contribution and Varian (2006) for a recent review.

[8] Taking preferences to be *total* answers directly to the claim that a single preference is not able to incorporate an array of different motivations, cares, or concerns including moral sentiments. But assuming that an agent is able to perform a total comparison is not innocuous. For instance, people may not be able to resolve the conflict between different or opposite concerns, cares, or motivations. These may not be commensurable in the sense that trade-offs are not possible between the different rankings. If this is the case, a complete ranking of alternatives may not be achieved.

[9] See Hausman (2012, 13-20) on how the concept of preference as a TSCE also relates and partly justifies the rationality of the standard assumptions of transitivity and completeness.

the alternatives they believe to be available, the economic agent will choose one that is at the top of their preference ranking. Still, Hausman (2012, 20) argues "[t]hat choices be determined by preferences is *not* demanded by rationality". For instance, it may be rational to use a heuristic to arrive to a choice (e.g. to avoid cognitive costs imposed by reason-based deliberation). According to Hausman (2012), rationality only demands that one *should* not choose x when y is available and one is confident that all-things considered y is preferred to x. But one can argue that intuitions, *feelings* (defined as emotional reactions), or *inclinations* (defined as idiosyncratic tastes) can provide rankings of alternatives that, in given choice situations, are all that is relevant for choice. Then, remark that if one "enlarges" the concept of preference beyond reason-based deliberation (e.g. to include preferences exclusively based on inclinations or feelings) preferences would *determine* choices more generally.

Second, Hausman (2012, 16-20) defends that preferences as TSCEs are "context-independent" departing from the idea that an alternative is supposed to specify everything "relevant" to preference (given the T in TSCE). The relevant characteristics, according to Hausman (2012), are given by whatever an individual takes into consideration in a rational deliberation on what she has most reason to do (given the E in TSCE). Hausman agrees with Broome's (1991, 103) view that "[o]utcomes should be distinguished as different if and only if they differ in a way that makes it *rational* to have a preference between them" [emphasis added]. Then, whether z (or something else deemed "irrelevant" to preference) is present should not matter for the preference between x and y. It follows, according to this view, that a TSCE (and the choices it determines) is *rational* if and only if it is context-independent. Thus, according to Hausman (2012), *rational choice* is the result of context-independent evaluations that provide reasons for choice. However, remark that it does not follow from these arguments that preferences are *stable* over time (as Hausman 2012, 16 recognizes). Take x and y to be two distinguishable alternatives for which time *per se* is irrelevant for the preference between them. Even if the preference between x and y is context-independent in Hausman's sense, this preference may change over time given changes in one's rational deliberation or, if one "enlarges" the concept of preference beyond reason-based deliberation, given potential changes in one's intuitions, feelings or inclinations.

Hausman (2012) gives four arguments in favor of sticking to his notion of preference. One of these, arguably the most important (see also Lehtinen 2012), is that, according to Hausman, only the conception of preference as a TSCE allows game theory and expected utility theory to serve their predictive and explanatory roles (see Hausman 2012, 65-70). Succinctly, the rationale of this advantage in terms of game theory is that if preferences are not considered as total comparisons then the game is incorrectly specified. That is, if economists do not include all the motivating factors into the payoffs of a game then the game does not correspond to the one that is actually being played. It follows that the analysis of such incompletely specified game will provide incorrect predictions or intuitions. Remark, once more, that such advantage would hold even if preferences are not seen as reason-based evaluations in the sense of Hausman (2012). What is needed is that preferences are total comparisons.

The other three arguments given by Hausman (2012, 64-5) are that a TSCE (i) matches economic practice, that (ii) it conforms roughly with the everyday usage of the word preference which helps avoid misunderstandings, and that (iii) it allows to pose questions concerning what preferences depend on. As noted by Lehtinen (2012), these arguments refer to pragmatic advantages, and for that reason carry low weight in deciding which notion of preference to adopt as a benchmark in economics. In my view, the third reason, although pragmatic in nature, is more relevant than the two former in that respect. In particular, it points out how sticking to Hausman's notion of preference as a benchmark for economic analysis may be useful to separate different notions and aims of research. In Hausman's (2012, 65) view, "[b]y treating preferences as total rankings, economists can separate the use of the word 'preference' from substantive views about what preferences depend on". As it becomes clear along the book, Hausman (2012) believes that economists should focus on how preferences are formed, and thinks that adopting a notion of preference as a TSCE is useful since, unlike some other notions of preference - such as preferences as exclusive expressions of tastes -, "it does not settle *a priori* what influences preferences" (p. 65). It suggest instead that preferences can be determined by several motivations. As before, this advantage seems to be shared by any definition of preference that is a total subjective comparison, but not necessarily a reason-based one.

A useful or misleading benchmark?

Hausman's (2012) arguments suggest that models based on the maximization of a single preference, when preference is seen as a stable TSCE, may be useful in terms of *parsimony*, *generality*, and *tractability*. In particular, such models allow economists to efficiently use the wide range of tools that they have developed so far, as it is the case of game theory. According to Hausman (2012, e.g. 73), this representation of human behavior and rationality is a useful benchmark when the main interest is the determination of action by the interplay of beliefs, constraints, and preferences. And, as it is the case with consumer choice theory, it seems that this is sometimes the case at least as a first approximation to the agents' choice behavior.

Still, seeing preferences as total comparative evaluations is a strong idealization. Besides excluding preferences that are not based on rational deliberation, Hausman's conception of preference is not compatible with several sources of preference change when preferences are taken to be stable.[10] In particular, and as developed below, a stable TSCE is not compatible with changes in preferences due to changes in values or other experiences of the agent. This is problematic, among other things, because the theory is silent with respect to the effect of changes in market rules or changes in other economic institutions on preferences *per se*.

The findings of psychology and behavioral economics also suggest that neither the choice behavior nor the decision making of most individuals accord with this theory. The accumulating evidence on the context-dependency of behavior and limits to rationality bring several doubts concerning the predictive capacity and normative value of the rational agent model. Next, I turn to one of the most prominent answers among behavioral economists that tries to address these issues but that ends up "replacing" the rational agent by an *inner or outer rational agent*[11].

3 The Inner and the Outer Rational Agents

Behavioral economics is by now a field interested in very different determinants of behavior, but the first and leading strand of literature has focused in inconsistent choices that result from intuitive or seemingly faulty psychological mechanisms (see Hoff and Stiglitz 2016). There is a long list of studies that show the effects of frames, anchoring, inattention, and other bias on decision making. In order to accommodate these findings, many authors have proposed incremental improvements to the standard rational choice model. In a review of many of these proposals, Rabin (2013, 528) argues that such improvements incorporate greater "realism while attempting to maintain the breadth of application, the precision of predictions, and the insights of neoclassical theory".

According to this view the maximization of a stable preference is seen as a useful benchmark or first approximation, from which behavioral theories are supposed to be judged against, for instance, with respect to their additional explanatory power. At the same time, many authors treat departures from the standard assumptions about rational choice as *mistakes* (e.g. Akerlof 1991; O'Donoghue and Rabin 2003; Bernheim and Rangel 2004). According to Rabin (2013, 529), "[w]e can capture many errors in terms of systematic mistakes in the proximate value function people maximize (quasi-maximization models)".

An example is provided by the literature on *preference reversals* in intertemporal choice between a smaller short-term reward and a larger long-term reward, that according to Rabin (2013, 534) is the most successful of the incremental improvements to neoclassical economics.[12] Many authors interpret these preferences as *present-biased*, and represent them in a two-parameter model that modifies exponential discounting (see e.g. Akerlof 1991; Laibson 1994, 1997; O'Donoghue and Rabin 1999, 2001, 2003). Let u_t be the instantaneous utility an individual derives from an activity in period $t \in \mathcal{T} = \{1, ..., T\}$. Then, these models (numerically) represent the individual's intertemporal preferences at period t with the following utility function:

$$\text{For all } t \in \mathcal{T} = \{1, ..., T\},$$

$$U^t(u_t, u_{t+1}, ..., u_T) \equiv \delta^t u_t + \beta \sum_{\tau=t+1}^{T} \delta^\tau u_\tau \tag{1}$$

[10] See e.g. Livet (2006) and Dietrich and List (2013, 2016) for discussions of this limitation and attempts to build theories of preference formation *and* preference change.

[11] I borrow the term "inner rational agent" from Infante et al. (2016). See their essay for a critical analysis of the approach discussed in the next section when related to welfare economics.

[12] In the typical example or experiment, agents choose between a smaller reward at period 2 and a larger reward at period 3. If the choice is made at period 2, then the smaller-earlier reward is chosen. If instead the choice is made prior to period 2, then the larger-later reward is chosen.

where $\beta > 0$ and $\delta \leq 1$. Remark that δ represents "time-consistent" discounting, and that if $\beta = 1$ these preferences represent standard time-consistent exponential discounting. Instead, if $\beta < 1$ these preferences are interpreted as "time-inconsistent" preferences for instantaneous utility (i.e., preferences that are present biased in the sense that the individual gives more relative weight to period τ in period τ than she does in any period prior to period τ). This present bias (and the preference reversal it entails) is often interpreted as a defect. For instance, O'Donoghue and Rabin (2003, 187) treat "this preference for immediate gratification as an error", and Rabin (2013, 538) regards the present bias to be a "quasi-maximization error".

Accordingly, some authors, such as Akerlof (1991), interpret an unobserved long-run intertemporal preferences (with $\beta = 1$, i.e., with the present bias "removed") as the "true" preferences of the individual. However, other authors such as O'Donoghue and Rabin (1999) do not assume that individuals have such "true" underlying preferences. Instead, they take the long-run intertemporal preferences with $\beta = 1$ to be "fictitious" and interpret it to represent the preferences that individuals *should* have would they have not been biased. O'Donoghue and Rabin (1999, 112-3) justify this assumption arguing that "[s]ince present-biased preferences are often meant to capture self-control problems, where people pursue immediate gratification on a day-to-day basis, we feel the natural perspective [for welfare analysis] in most situations is the 'long-run perspective'." What is assumed is that the *best* for individuals would have been to follow the reasoning of a rational agent, that they have not followed because they are psychologically bias or *naive*. Then, these authors associate normative authority to a given and unobserved preference that is time- and context-independent and implicitly assume that any deviation from this mode of reasoning is a mistake.

Since committing a mistake is, by definition, making an action that departs from something that is true, proper, or right, models such as O'Donoghue and Rabin (1999) implicitly assume that the true, proper, or right thing to do is to maximize an unobserved and stable preference relation. This benchmark is taken as the right thing to do even if it contradicts the preferences revealed by the individual. Moreover, contrary to Akerlof (1991), O'Donoghue and Rabin (1999) do not associate the unobserved rational preference with an underlying "true" preference of the individual. Then, models in this vein can be seen as taking an *outer rational agent* (i.e., a fully rational agent that is not part of the individual and may disagree with the individual's preferences), as the guide for *sophisticated* behavior and the source of normative authority.

Consider now another interpretation of the problem of preference reversals, that represents the intrapersonal conflict between present and future preferences with "dual-self" or "dual-mode" models (e.g. Thaler and Shefrin 1981; Bernheim and Rangel 2004; Fudenberg and Levine 2006, 2012). These models represent intertemporal decisions as intrapersonal interactions between two selves or two modes: one impulsive and myopic and the other patient and far-sighted. For instance, Fudenberg and Levine (2006) treat the two selves as two *rational* players in an intrapersonal game: a "planner" (concerned with lifetime consumption) and a "doer" (that exists only for one period and is only interested in the consumption of that period). Bernheim and Rangel (2004) build a similar model to study addictive behavior, in which according to the exposure to "environmental cues", an individual alternates between a "hot mode" in which she always takes an addictive behavior "irrespective of underlying preferences", and a "cold mode" in which she "considers all alternatives and contemplates all consequences" and selects her most preferred alternative. In this sense, Bernheim and Rangel (2004) assume that each individual has a stable and single preference relation, and that choices taken under the hot mode are "mistakes" that may differ from the choices that would be determined by the individual's preferences. Instead, Fudenberg and Levine (2006) assume that both players are rational maximizers who have the same short-run preferences, and interpret the individuals' seemingly mistakes as cases of high self-control costs. But in both cases, the long-run perspective (the planner or the cold mode) is taken as the source of normative authority.

These models are related to the two selves model recently popularized by Daniel Kahneman (2011). According to the author, human psychology can be divided into two "systems" or modes of thought: one fast, effortless, and automatic (System 1), and another slow, effortful, and controlled (System 2). These two systems correspond "roughly to intuition and reasoning", and while System 1 generates involuntary impressions of the objects of perceptions and thought, System 2 is involved in all intentional judgments based on impressions or deliberate reasoning (see also Kahneman 2003). System 2 is also thought to monitor the activities of System 1, and the preferences of the former are not necessarily consistent with the preferences of the later. In this light, the "doer" in Fudenberg and Levine (2006) and the "hot mode" in Bernheim and Rangel (2004) are separate systems responsible for intuitive choices, while the respective "planner" and "cold mode" are separate systems responsible for reason-based actions.

As argued by Infante et al. (2016), these models suggest that individuals are endowed with an *inner rational agent*, i.e., an independent agent that is able to make stable and context-independent decisions *exclusively* based on its own evaluations of alternatives.[13] In this view, "human psychology is represented as a set of forces which affects behaviour by *interfering with* rational choice", that is itself "represented by the error-free reasoning of the inner agent" (Infante et al. 2016, 14-5). Decisions based on a psychological bias or a System 1 (hereafter intuition) are in this perspective deemed not to reveal an authentic and/or normatively relevant preference, and the agent is often assumed to have a stable and context-independent preference based on a slow, deliberated, and controlled System 2 (hereafter reasoning).

A useful or misleading benchmark?

I have distinguished two stances. One does not assume that agents are endowed with an underlying system capable of rational decision making, while the other assumes that such an inner rational agent exists. At the same time, we have seen two common trends: (i) to treat choices that result from psychological bias or intuition as mistakes (even though some authors try to avoid this assumption as e.g. Fudenberg and Levine 2006), and (ii) to assume that choices that result from psychological bias or intuition do *not* reveal authentic and/or normatively relevant preferences. The behavior of an inner or outer rational agent is taken as the source of normative authority, and the traditional rational agent model is used as a reference point or approximation from which extensions and variations are found and modeled.

Building models that are incremental improvements of the rational agent model is often a useful strategy in economics. Besides favoring parsimony, tractability, and generality, it allows to observationally distinguish between choices that result from a stable preference and choices that result from other factors. Models in this vein may be particularly useful in cases of choices that are uncontroversially determined by other factors than what can reasonably be associated with an individual's preferences. Bernheim and Rangel (2004, 1561-2) give the example of American visitors to the United Kingdom who suffer injuries and fatalities because they only look to the left before crossing the street, even though they know that traffic approaches from the right. It seems clear that one "cannot reasonably attribute this to the pleasure of looking left or to masochistic preferences". The behavior, in this case, can be uncontroversially considered as resulting from a mistake.

However, in the remainder of the paper I wish to argue that interpreting any deviation of "rational" behavior as mistakes is often an overly "mechanical" recipe for both positive and normative economics. Namely, I will argue that recognizing that not all these decisions are the result of mistakes, and that individuals themselves are, *a priori*, the best (or at least the first) judges about the nature of a decision, may help economists to design better explanations and predictions of behavior as well as better welfare criteria.

4 Mistakes or Reflexive Decisions?

> "May I urge that changes in values do occur from time to time in the lives of individuals, of generations, and from one generation to another, and that those changes and their effects on behavior are worth exploring - that, in brief, de valoribus *est* disputandum?" (Hirschman 1984, 90)

The previous two Sections raise important questions. Hausman's (2012) view in favor of a conception of preference as the result of rational deliberation and the models that treat either psychological bias or intuition as mistaken raise the following questions: (i) can and ought a choice based on intuition (be interpreted to) reveal a preference, and (ii) can and ought a choice based on a psychological bias or intuition *not* always be (interpreted as) a mistake. As for the first, recall that according to Hausman's (2012) definition of preference choices based on intuition cannot reveal a preference because such choices are not evaluations based on a *rational deliberation* about what the individual has most reason to do. However, intuitions, like feelings or inclinations, seem to be important components of individuals' evaluations of alternatives that, in some occasions, are part of a person's rational deliberation about her reasons to choose. Then, it seems strange to disregard these motivations - intuitions, feelings, or intuitions - as determinants of preferences in the occasions that they determine choices even though those choices are not the result

[13] These are the models, as pointed by Infante et al. (2016), that are the closest to explain this assumption. See their essay for a critic of other models that implicitly assume the existence of an inner rational agent without specifying an underlying model of rationality.

of rational deliberation (specially in a positive theory of behavior). As argued above, allowing preference rankings to be also determined exclusively by intuitions, feelings, or inclinations is more consistent with the assumption of choice determination and would be consistent with context-independent but unstable preferences.

Moreover, upon rational deliberation I may evaluate two alternatives to be equally worthy in terms of all reasons besides an intuition, feeling, or inclination in favor of one of the two alternatives. Then, I may *prefer x to y* - according to Hausman's (2012) definition - based on a first automatic impression (intuition), feeling, or inclination. Although a choice based on this preference is not an example of a "fast" choice exclusively based on intuition (since it is made after rational deliberation), it illustrates how it is possible to reveal a preference - in Hausman's (2012) sense - for one alternative over another because of a first automatic impression.[14]

One could argue that the choice just described reflects one's natural tendencies or inclinations, instead of the expression of an "authentic" preference between the two alternatives. A similar argument could be made for choices exclusively based on intuition, feelings, or inclinations. Although I favor calling such tendencies and inclinations expressions of preference, I think such argument points towards the relevance of the following distinction: between the preferences/inclinations that a person identifies with and the ones that a person does not identify with. If, say, upon reflection, I identify with the choice of x over y that I made based on intuition, this choice, even if not based upon a process of slow rational deliberation, seems to have revealed a preference for x over y that I (and not an observer) deem authentic. I call such "self-authenticated" preference a *reflexive preference*.

According to this view, choices can reveal preferences when based on either reasoning or intuition, but it is important to distinguish between preferences with which people identify and preferences with which people do not identify (reflexive and non-reflexive preferences respectively).[15] This distinction can be conceptualized through people's preferences over preferences (or choices), also known as hierarchical preferences, meta-preferences, or second-order preferences. For example, "I would prefer not to prefer to smoke" (to non-smoking) is a second-order preference. And since preferences in general determine choices in economics, it is often possible to describe such preferences in relation to observed choices such as "I would prefer myself not to smoke". In the next Section I "decompose" second-order preferences into two independent rankings, but before that I wish to discuss some of their features and their relation to the notions of mistake and preference change.

Three features of second-order preferences are worth mentioning. The first is that second-order preferences correspond to an important part of people's *values*[16] (see also Hirschman 1984). This is the view of Lewis (1989, 115), who argues that desiring to desire (a second-order desire) is *valuing*:

> "The thoughtful addict may desire his euphoric daze, but not value it. Even apart from all the costs and risks, he may hate himself for desiring something he values not at all. It is a desire he wants very much to be rid of. He desires his high, but he does not desire to desire it, and in fact he desires not to desire it. He does not desire an unaltered, mundane state of consciousness, but he does desire to desire it. We conclude that he does not value what he desires, but rather he values what he desires to desire."

The second feature is that second-order preferences are not only relevant for cases of (lack of) self-control. In many applications, the conflict between first- and second-order preferences is indeed related with addictive or impulsive behavior that leads to a lack of self-control (e.g. smoking, drugs use, betrayal). Second-order preferences are a way to rationalize and predict such behavior that in general escapes the traditional rational choice model. However, second-order preferences are also relevant to inform cases related with the values, goals, and aspirations of individuals that are not related with self-control. For example, I may want myself to prefer to do more non-paid voluntary work, but have a first-order preference for more paid work given budget constraints.

[14] Infante et al. (2016) build a related but different argument based on the possibility that, all-things considered, two alternatives may be incomparable. The authors argue that in that case it is rational to choose based on an inclination, and that for that reason the individual's choices may be context-dependent. One question that emerges is if this inclination can be separated, as Infante et al. (2016) seem to assume, from the all-things considered rational deliberation.

[15] Examples of choices that, according to this view, do not reveal preferences are choices based on manipulation. I discuss issues related with adaptation and false beliefs in Section 8.3.

[16] Values are here defined as *something* (in this case preferences) intrinsically valuable or desirable.

The third feature is that second-order preferences are not *direct* determinants of choices, i.e., they do not necessarily imply action.[17] For instance, many heroin addicts, even if they do not identify with their preference for heroin and would like to quit using drugs, will often fail to do it. Even in many non-addictive behaviors of our day-to-day life, we often behave in ways that we do not identify with and/or that do not accord with our values. Second-order preferences are likely to be taken into consideration in an all-things considered rational deliberation, but they do not directly imply that preferences or action will be aligned with them.

I would like now to argue that a choice resulting from a psychological bias or intuition should not always be interpreted as a mistake. As remarked before, committing a mistake is, by definition, making an action that departs from something that is true, proper, or right. A sensible, if not natural, reference of truthfulness for a person is *who* this person is and what she values. According to this view, a person's choice is a mistake if the choice does not correspond to *who* she is and/or what she values. In most cases, the best judge of who one is the individual herself. The individual is also the most likely to know her values.

My suggestion is that a mistake is an action that I judge as mistaken.[18] It follows that a choice based on a psychological bias or intuition *may not* be a mistake. For instance, I may recognize that I acted based on intuition when I bought that delicious ice-cream that ruined my diet. However, in retrospect, I may not consider it as a mistake: the diet was not reflecting, according to my own judgment, who I am, want to be or become.

Similar considerations apply to the example of preference reversals in intertemporal choice that we have seen in the previous Section. It may be the case that a person does not judge her present-bias as a mistake, but instead see it positively and identifies with it. For instance, it may be the case that the person wants to live her life at the fullest and values (has a second-order preference for) higher immediate gratification against lower future utility. This example illustrates that it is different to assume *a-priori* that a present bias is a defect, then to assess if this is indeed the case according to the person's own judgments about her preferences. In this sense, second-order preferences tell us, from the individual's own perspective, if she endorses or not the present bias.

To sum up, I have argued that a choice *exclusively* based on intuition can reveal an authentic preference, and that not all choices based on a psychological bias or intuition should be treated as mistakes. I have argued that what is important is to distinguish between "self-authenticated" preferences and preferences that individuals do not identify with. It results that choices that violate the traditional axioms of revealed preference and that are often treated as mistakes may be instead the result of preference change that individuals identify with. I call such changes *reflexive preference changes.*

Reflexive preference changes can be conceptualized as preference changes that result from the resolution of a conflict between a first-order preference and a second-order preference (see also Hirschman 1984). For instance, I may have a first-order preference for eating meat but form a second-order preference that values it negatively because of ethical, environmental, or other reasons and become a vegetarian such that I align my preferences with my values.

There are several reasons why it is important to consider reflexive preferences and reflexive preference change in economics. First, recognizing that changes in preferences may be the result of reflective decisions may lead to better description and prediction of economic behavior (see also Hirschman 1984, 90). For example, distinguishing between preference change due to changes in values and preference change due to changes in tastes is important since values and tastes are not, *a priori*, susceptible to be affected similarly by changes in market rules or other economic institutions. Similarly, distinguishing choices that individuals identify with and self-authenticated mistakes may help the correct interpretation of data from the "real world" or from laboratory experiments. Second, reflexive preferences and reflexive preference change can describe and predict phenomenons of interest to economics that are not captured by the traditional rational choice model (such as lack of self-control). Finally, these notions are relevant for welfare inference and policy analysis. For example, knowing if "inconsistent" choices are the result of self-authenticated mistakes or choices that the individuals identify with may help to refine behavioral welfare rankings proposed in the literature (see Section 7).

[17] According to Lewis (1989), you are disposed to follow your second-order preferences/values if you were put under hypothetical "ideal conditions" to follow them. This means that, according to the author, second-order preferences *directly* determine choices only under conditions that are not usually met in real life. See also Frankfurt (1988).

[18] This *self-authenticated* definition contrasts with an *objective* definition of mistakes that may include choices based on adaptation or false beliefs as seen from the point of view of the observer. I discuss these issues in Section 8.3.

As argued by Hirschman (1984, 90), the possibility of reflexive preference change brings an important argument against the view, defended by Stigler and Becker (1977), that all economic behavior change should be explained and understood through changes in prices and incomes. In fact, Stigler and Becker (1977) argued that preference change is of little interest since it often results from "capricious" changes in tastes.[19] By contrast, reflexive preference change suggests a non-capricious way by which change in preferences may occur and that, according to the view defended here, economists should not overlook.

5 Hierarchical Models

In what follows I sketch two *hierarchical* models. The hierarchical *retrospective model* takes a backward-looking perspective, where choices are judged by the individual *ex-post*. This model is, among other things, adapted for welfare analysis. When presenting it, I introduce some formal definitions that will be used later in the refinement of a behavioral welfare ranking (Section 7.2). In particular, I "decompose" second-order preferences into two independent rankings: identification *and* valuation preferences. The hierarchical *evolving model* takes a forward-looking perspective, where choices are the result of the conflict (or resolution) between first- and second-order preferences. I informally explore how this model could explain preference and choice reversals. In the next Section, I discuss two models of the economic agent that could be used in combination with a hierarchical evolving model to formalize these notions in the future. Before proceeding, I introduce some general definitions and the framework of choice with time that will be used[20]

5.1 Framework

Let X be a universe of alternatives that are of interest to the economic agent. Different alternatives, denoted by x, y, etc., can be standard objects such as consumption bundles, but alternatives can also include non-standard features as long as they are complete and mutually exclusive descriptions of the world. For example, an alternative may include the description of possible actions (instead of standard objects) when combined with the remaining description of the world. Let $\mathcal{P}(X)$ be the set of all non-empty subsets of X, and $\mathcal{F} \subseteq \mathcal{P}(X)$ be a collection of subsets of X. Each of these subsets is interpreted as a *choice problem* (using Arrow's 1959 terminology), sometimes called a choice situation.

In what follows, I will often refer to two-element subsets of X (hereafter binary choices). Binary choices are related to many of the examples I have provided, and are intimately linked to the concept of rational choice.[21] They also illustrate the possibility of choosing between action and inaction (e.g. between "smoking" and "not smoking" when the two alternatives differ only in this respect).

In the following, I adopt the framework that is developed in Chapter 3. Let $\mathcal{T} = \{1, ..., T\}$ denote a discrete time horizon, and $K : \mathcal{T} \longrightarrow \mathcal{F}$ a *chronology of choices* that assigns to every choice period $t \in \mathcal{T}$ a unique non-empty set $A(t)$, interpreted as the choice problem taking place at *period t*. Note that K may be any sequence of choice problems, and include the same non-empty set at two distinct choice periods. As in Chapter 3, a *chronological choice function* C is a mapping that assigns to every pair $(t, A(t))$ of a chronology K a unique element $C(t, A(t)) \in A(t)$. $C(t, A(t))$ is interpreted as the chosen alternative at $(t, A(t))$.

As in the standard theory of choice, it is possible to define what it means for an agent to be *consistent* in her choices. One definition of consistent choice, denoted $x \succsim^C y$, can be stated as follows:

Definition 1 Alternative x is said to be **consistently** chosen over alternative y (or $x \succsim^C y$) if and only if $y \neq C(t, A(t))$ for all $A(t) \in X$ such that $x, y \in A(t)$ and $x = C(t, A(t))$ for some $A(t) \in X$ such that $x, y \in A(t)$.

In words, one alternative is consistently chosen over another if the former alternative has been chosen at least once when the latter was present and the latter has never been chosen when the former was present.

[19] Stigler and Becker (1977, 89) also argued that changing preferences provide "endless degrees of freedom". See Fehr and Hoff (2011, 398-400) for why this is not a substantive argument with today's knowledge about preference explanations.

[20] I adopt a framework of choice with time that is first developed in a joint work with Nicolas Gravel and presented later in Chapter 3 of this dissertation.

[21] See Bossert et al. (2005, 2006) for recent work on the rationalizability of choice functions on the domain that includes all singletons and all two-element subsets of X.

5.2 Hierarchical Retrospective Model

The *retrospective model* is defined over observed choices. I interpret choices to reveal the agent's ***actual preferences***[22]. Remark that actual preferences may not be the result of rational deliberation as in Hausman's (2012) sense, i.e., they may be determined by desires based on inclinations, feelings, or intuitions. In this model the reflexive attitudes take place in retrospect at period T (the end of the time horizon). This seems to be a reasonable assumption if we are interested, for example, in determining which preferences/choices to take into account in terms of individual well-being *ex-post* (see Section 7). It is less appealing, for example, if one is interested in explaining preference change or for the prediction of behavior from one period to another (see Section 5.3).

For all $(t, A(t))$, define $\succsim_t^I$ on any $A(t)$. For any $x, y \in A(t)$, $x \succ_t^I y$ if and only if the agent *identifies with the choice of* x *from* $A(t)$ *more than with the choice of* y *from* $A(t)$.[23] Judgments of the type $\succsim_t^I$ are interpreted as the agent's ***identification preferences*** at period T. They can be revealed, for example, through stated second-order attitudes such as "I would myself want to have chosen x from $A(t)$ more than have chosen y from $A(t)$". The preference $\succsim_t^I$ is not necessarily complete and transitive, but is assumed to be reflexive (as a binary relation) and acyclic. I now define what I mean by a reflexive choice:

Definition 2 A choice $C(t, A(t))$ is said to be **reflexive** if and only if $C(t, A(t)) \succsim_t^I x$ for some distinct $x \in A(t)$ and $y \succ_t^I C(t, A(t))$ for no distinct $y \in A(t)$.

In words, a choice is said to be reflexive if in retrospect the agent weakly identifies with the chosen alternative with respect to at least one other feasible alternative and for no other feasible alternative it is the case that she identifies more with that alternative than with the chosen one. In the case of a binary choice, a choice is said to be reflexive if in retrospect the agent weakly identifies with the chosen alternative with respect to the other feasible alternative. Then, for any binary choice, an agent may either identify with the chosen alternative more than with the other alternative, be indifferent in terms of identification between the two alternatives, do not identify with the chosen alternative in the sense that she identifies more with the other possible alternative, or have no identification preference between the two alternatives.

Note that one could certainly weaken or strengthen the definition of a reflexive choice. For example, one could require that the chosen alternative is strictly preferred in terms of identification to at least one alternative and/or exclude the possibility of indifference in terms of identification. Note also that I abstract from some important issues with this formulation. For example, it is possible that an agent identifies more with a chosen alternative than with another feasible alternative but not *enough* for her to identify with the choice itself. It is also possible that an agent identifies with a choice itself, though there is no feasible alternative for which she identifies less than the chosen one. But for the current purposes, I stick with this definition.

Clearly, a choice may not be reflexive in this sense. For example, a choice (or the actual preference behind it) may not agree with the identification preference over it. Whenever an agent makes a choice that is against her identification preferences, then this choice seems to be judged negatively *by the agent herself.* Then:

Definition 3 A choice $C(t, A(t))$ is said to be a (self-authenticated) **mistake** if and only if there exists a distinct alternative $x \in A(t)$ such that $x \succ_t^I C(t, A(t))$.

In other words, a choice is said to be a mistake if there exists at least one feasible alternative that the agent identifies more than with the chosen one. For the case of binary choices, this is the converse of a reflexive choice. In case of subsets with more than two alternatives, it is possible to have a choice that is neither reflexive nor a mistake: it suffices that $C(t, A(t)) \succsim_t^I x$ for no distinct $x \in A(t)$ and $y \succ_t^I C(t, A(t))$ for no distinct $y \in A(t)$.

I now relate Definition 1 with Definition 2. If between periods 1 and T the agent makes a choice over two alternatives based exclusively on reflexive preferences, then one can say that she identifies with *all* her choices over these two alternatives. This lead us to the following definition:

[22] I borrow this term from Harsanyi (1997). See Section 8.3 below for the contrast with *informed preferences.*

[23] Remark that these preferences, though formally defined as first-order, are interpreted to be of second-order because they are over alternatives contained in past choice problems where it is assumed that choices reveal actual preferences. In this sense, $x \succ_t^I y$ can be read as if $(x \succ_t y) \succ_t^I (y \succ_t x)$. See Watson (1975, 219) for an alternative view to Frankfurt (1971) based on the perspective that some second-order attitudes can be instead seen as first-order.

Definition 4 Alternative x is said to be **reflexive-consistently chosen** over alternative y if and only if $x \succsim^C y$ and for all $A(t) \in X$ such that $x, y \in A(t)$ and $x = C(t, A(t))$ it is the case that $x \succsim_t^I w$ for some distinct $w \in A(t)$ and $z \succ_t^I x$ for no distinct $z \in A(t)$.

In words, an alternative x is said to be consistently chosen in a reflexive fashion over an alternative y if x is consistently chosen over y and whenever x is chosen and y is present the choice is reflexive. In the case of binary choices, this definition is independent with respect to other alternatives when judging if the choices between two alternatives have been reflexively consistent. Otherwise it is not. An alternative definition could impose a certain independence with respect to other alternatives for any choice problem.

Besides identification preferences, the agent is assumed to have reflexive and acyclic (possibly incomplete and not necessarily transitive) preferences over her different choices. This means that the agent is able to make statements of the sort $C(t, A(t)) \succsim^V C(t, A(t'))$, where $\succsim^V$ is defined over $\cup_{t \in \mathcal{T}} C(t, A(t)) \subseteq X$. These statements are interpreted as her ***valuation preferences***, and reflect her evaluative judgments about the relative importance of different choices (or what the agent cares about).[24] These preferences incorporate a sense of valuation that is often absent in most utility views of agency and well-being. As argued by Sen (1987, 19-20), "valuation is a *reflexive* activity in a way that 'being happy' or 'desiring' need not be". As it was the case with identification preferences, valuation preferences are taken to be the result of the agent's reflexive activity at period T. In some examples and one of the applications discussed below these preferences will not be used (Section 7.2). But in general, if they exist, they seem to provide valuable information on what people take to be of value or care about.

The distinction of the two types of preferences seems to be descriptively meaningful. In particular, they seem to reflect preferential (reflexive) judgments of different natures: the former concerning the identification with the alternatives that could have been chosen in each choice problem, and the latter concerning a value ranking over different choices or actual preferences. Moreover, identification preferences can relate an actual choice (e.g. "smoking" has been actually chosen over "not smoking") to an hypothetical choice ("not smoking" being chosen over "smoking", even though this has never been observed as a choice); on the contrary, valuation preferences are defined over the several actual choices that have been observed.

Note that these rankings differ from the traditional view of individual meta-preferences adopted in economics.[25] In general, meta-preferences are assumed to be an unique ordering over (hypothetical) multiple first-order preferences over the universe of alternatives. In this Chapter, I interpreted two independent rankings as reflecting the agent's second-order retrospective preferences over her past preferences or choices. These rankings also differ from other meta-rankings based on morality, ideology, or political priorities, in that identification and valuation preferences are based on *individual* evaluations instead of an observer's point of view.

5.3 Hierarchical Evolving Model

Contrary to the retrospective model, in the evolving model the reflexive attitudes take place at every $t - \epsilon$ with $\epsilon \in]0, 1[$. In other words, the agent is assumed to judge the preferences over her potential choices before every period. This formulation could capture some of the essential features of *second-order volitions* (i.e., second-order desires that first-order desires effectively motivate or move you to action) that are at the heart of the philosophical thought about second-order attitudes, and that are absent from a hierarchical retrospective model. Bratman (2003, 224) summarizes the features that are common to many (hierarchical) models that consider second-order volitions in the tradition of Frankfurt (1971):

> "First, it will involve a second-order attitude that is about that desire. Second, this second-order attitude will itself be a conative attitude, in the broad, generic sense of a motivating attitude. Third, this second-order conative attitude will concern certain kinds of further functioning, from now on, of the first-order desire. The content of this second-order attitude will be in this sense forward-looking. Fourth, this forward-looking second-order conative attitude will include in its own functioning the guidance, from now on, of the functioning of the first-order desire. In short:

[24] See Decancq et al. (2015) for a similar definition over functionings instead of choices/preferences.

[25] See Sen (1977) for a brief discussion of different interpretations of meta-rankings. Although Sen (1977) focus on an observer's "moral" ranking, he notices that a meta-ranking can "be ordered also on grounds other than a particular system of morality", such as the "preferences one would have preferred to have" (1977, 338-9).

> the theory will appeal to a higher-order attitude that is conative, forward-looking in its content, and guiding in its function.
> There is also a fifth feature that such theories try to capture. The higher-order, forward-looking, and guiding conative attitude is to constitute - at least in part, and given relevant background conditions - a commitment on the part of the agent concerning the role of the target desire in her own agency: the agent is appropriately settled on this."

In what follows, I sketch analogous definitions to the ones of the retrospective model. These could be used for the *ex-post* evaluation of choices, but also to predict future behavior. For all $(t, A(t))$, let $\succsim_t^{I_{t-\epsilon}}$ be defined on any $A(t)$. For any $x, y \in A(t)$, $x \succ_t^{I_{t-\epsilon}} y$ if and only if the agent *identifies with the choice of x from $A(t)$ more than with the choice of y from $A(t)$ at period $t-\epsilon$*. Judgments of the type $\succsim_t^{I_{t-\epsilon}}$ are interpreted as the agent's **identification volitions** at period $t-\epsilon$, i.e., conative, forward-looking, guiding and committed attitudes in favor of choosing one alternative over others at period t. The preference $\succsim_t^{I_{t-\epsilon}}$ is again assumed to be reflexive and acyclic, but not necessarily complete or transitive. They can be recovered, for example, through stated second-order attitudes at period $t-\epsilon$ such as "I would myself want to choose x from $A(t)$ more than choose y from $A(t)$". Then:

Definition 5 A choice $C(t, A(t))$ is said to be **ϵ-reflexive** if and only if $C(t, A(t)) \succsim_t^{I_{t-\epsilon}} x$ for some distinct $x \in A(t)$ and $y \succ_t^{I_{t-\epsilon}} C(t, A(t))$ for no distinct $y \in A(t)$.

Definition 6 A choice $C(t, A(t))$ is said to be a (self-authenticated) **ϵ-mistake** if and only if there exists a distinct alternative $x \in A(t)$ such that $x \succ_t^{I_{t-\epsilon}} C(t, A(t))$.

Definition 7 Alternative x is said to be **ϵ-reflexive-consistently chosen** over alternative y if and only if $x \succsim^C y$ and for all $A(t) \in X$ such that $x, y \in A(t)$ and $x = C(t, A(t))$ it is the case that $x \succsim_t^{I_{t-\epsilon}} w$ for some distinct $w \in A(t)$ and $z \succ_t^{I_{t-\epsilon}} x$ for no distinct $z \in A(t)$.

It is worth emphasizing that empirically, recovering this type of data is much more demanding than from a retrospective perspective. In particular, while the retrospective model involves the elicitation of preferences at one period of time, the evolving model entails the elicitation of identification volitions before each period.

Note that the notion of second-order volitions corresponds, in its most basic sense, to a theory of preference formation and change. It would be then possible to impose an axiomatic structure that would require that choices are the result of preferences motivated by second-order volitions. For example, whenever these volitions would move an agent towards a change in preferences, an evolving model could require such change in preferences to occur and the potential subsequent choice reversal to follow. This would be in line with the economic tradition of identifying the observable choice implications of different decision making models, as it is explored in Chapter 3 of this dissertation. It is worth noticing, however, that such formulation would be somewhat contrary to the philosophical view of second-order volitions. Even though their feature of commitment, an agent may fail to follow the decision he has taken in accordance to a second-order volition due, for instance, to a strong and irresistible desire to choose otherwise (see e.g. Frankfurt 1988). A non-deterministic view of agency, as alluded in the conclusion of this Chapter, could potentially explain and rationalize this kind of behavior. In any case, the development of such formal framework is a possible venture left for future work.

6 The Reflexive Agent

The possibility of reflexive preference change suggests the relevance of models that incorporate and explain changing (first-order) preferences. Decision making models based on multiple preferences are a possibility. They have shown to be a successful instrument to provide rationalizations for changing preferences and choice heuristics that may be behind some of the cyclical patterns of choice observed in real data (e.g. Aizerman and Malishevski 1981; Manzini and Mariotti 2007; see also the changing preferences model of Chapter 3). One way to represent a reflexive agent is then through multiple preferences that she alternates over time *and* identify or not with. This presupposes combining multiple preferences models with an explanation of reflexive preference change, such as the hierarchical evolving model just described.

Observationally, while the multiple (first-order) preferences may be recovered through choice data, data on second-order preferences may be collected, notably, from non-choice data such as individuals' verbal evaluations of their preferences/choices. In Section 8.1 I return to this topic and briefly discuss one survey-based and one choice-based method of eliciting second-order preferences.

In what follows, I discuss two representations of the economic agent that can be associated with an hierarchical (evolving) model in order to model reflexive and non-reflexive preference change. One is based on the conflict between an individual's multiple preferences (Section 6.1). The other is based on the evolution of an individual's single preference (Section 6.2).

6.1 The Conflicted Agent

> "People behave sometimes as if they had two selves, one who wants clean lungs and a long life and another who adores tobacco, or one who wants to improve himself by reading Adam Smith's theory of self-command (in *The Theory of Moral Sentiments*) and another who would rather watch an old movie on television." (Schelling 1984, 58)

The conflict between different preferences is a topic that has been widely discussed and modeled in economics. Models in this tradition represent the economic agent as if endowed with a collection of preferences, i.e., a plurality of distinguishable identities, roles, motivations, or points of view that are fixed over time. In some cases these are represented as a collection of orderings, and others as multiple *selves* (or "subagents") that interact with each other as if they were players in an interpersonal game. This representation, as noted by Gul and Pesendorfer (2008, 30), represents a "departure from the standard economics conception of the individual as the unit of agency".

This representation is in tune with views such that of Schelling (1984), who considers that people are best represented as a collection of "values centers" that share the same beliefs and reasoning capacities but differ in terms of volitions. According to his view, one value center (or self) will act as if a dictator at each period, winning "the intimate contest for self-command" at that period (see Schelling 1984, 57-81). From period to period, individuals "alternate" from one preference to another. An alternative and recent example is given by the *reason-based* theory developed by Dietrich and List (2013, 2016), in which an agent's preferences over alternatives depend on her *motivational state*, defined as a subset of all possible *motivationally salient* properties of those alternatives (the properties that the agent focus on). Then, in their model, an agent is represented as a family of preference relations over all possible motivational states and alternates from one preference relation to another according to the motivational state in which she happens to be.

Kalai et al. (2002) explore a testable model consistent with this view, in which an agent chooses the best alternative according to one of multiple preferences in each choice situation. Choice behavior is rationalized by a collection of preference relations, such that for every choice situation A in the domain of feasible choice situations, the chosen alternative is maximal in A for some preference in the collection. Note that with such a decision making model, one can always rationalize any choice behavior whatsoever by resorting to a sufficiently large collection of different preferences. Still, Kalai et al. (2002) show that a plausible upper bound on the number of different preferences that can rationalize a choice behavior generated by a universe containing n alternatives is $n-1$. More recently, Apesteguia and Ballester (2010) have built upon this framework and studied the complexity of finding the minimal number of preference relations - the lower bound - necessary to explain choice behavior.

Conflicted agent models could possibly explain the preference changes/reversals discussed earlier. If an individual identifies with her multiple preferences, and alternates between those she identifies when choosing, then the behavior of this person may be inconsistent with the standard revealed preference axioms without being interpreted as a mistake. If, instead, the individual does not identify with some of her multiple preferences, then some changes in the preferences exhibited by the choices of this person may not be the result of reflexive preference change. To distinguish reflexive from non-reflexive preference change, conflicted agent models need be coupled with an hierarchical model in order to specify if, at any given point in time, the multiple preferences and identification preferences of the agent are aligned.

In addition, conflicted agent models may be reasonable approximations of the mode of *reasoning* of individuals, or, at least, a convenient way of describing human psychology. One important distinction is between models that assume that the conflict is the result of the switch among a collection of preferences (as e.g. Kalai et al. 2002 and Dietrich and List 2013, 2016), and the models that assume that the conflict is the result of the strategic interaction between multiple selves modeled as "subagents" (as e.g. Schelling 1984). As for the former, they seem useful to study the behavior of individuals that switch preferences according to the *role* they happen to be playing or according to the (social) identity or motivation that happens to be salient at a given point in time. Recent experiments suggest that this may be a meaningful exercise. For example, in the case of Asian-American subjects, some experiments suggest that it is sufficient to make one preference (identity) more salient than another (the Asian or the American identity) to trigger different behavioral responses in terms of patience (Benjamin et al. 2010) or cooperation (LeBoeuf et al. 2010).

As for the later, the analogy between interpersonal and intrapersonal conflict is useful since economists have developed a wide range of tools to study interpersonal conflict that can, in this way, be used to study cases of intrapersonal conflict. However, it is worth noticing that the analogy between intrapersonal and interpersonal conflict may be sometimes misleading. For instance, "[p]eople are able to punish or control each other to avoid conflict in a way that is not possible among "multiple selves"" (Arlegi and Teschl 2015). According to Elster (1986, 2) "the possibility of mutual strategic interaction [...] is hardly plausible", and *deception* and *manipulation* may be the essential forms of interaction. And despite neurological evidence that the brain may sometimes work in this fashion (see e.g. Jamison and Wegener 2010), it seems that we still need to understand when such representation is an accurate description of *why* people make their decisions. Only with correct underlying assumptions about the individuals' motivations and mode of reasoning are we able to fully understand cause-and-effect relationships, and make predictions that will remain correct throughout different environments (see Schotter 2008, 72).

Finally, one can remark that the change in preferences in a conflicted agent model is made from a *fixed* collection of preferences. In a dynamic perspective, this means that the possible preferences that the agent can hold do not change over time. This is unlikely to hold in cases in which a person's values/second-order preferences change over time. Next, I discuss a representation of the economic agent that is in tune with the endogenous change of preferences as the ability to evaluate and change one's single preference, one's identity, or *who* one is or want to become.

6.2 The Evolving Agent

> "Thus we say of an oak that it is the same thing from the seed to the tree in the prime of life. The same is true for an animal from birth to death, and for a man, as a specimen of the species, from foetus to old man. The demonstration of this continuity functions as a criterion supplementary to that of similarity in the service of numerical identity. The contrary of identity taken in this third sense is *discontinuity.* But what has to be taken into account in this third sense is change through time." (Ricoeur, 1991, 190)

An alternative representation of the economic agent is to assume that the agent is endowed with one personal identity that evolves over time. This conceptualizes the individual as an *evolving agent* that makes her decisions according to an endogenous sequence of (multiple) preferences. Contrary to the rational agent model that assumes a stable personal identity over time, and differently than the conflicted agent model that assumes independent and conflicting preferences or selves at each period, the evolving agent model is a representation of the economic agent with a single preference that evolves over time.

While in economics, to the best of my knowledge, the endogenous evolution of a single preference is not a representation that is often used, the evolving agent model is consistent with a notion of the self (or personal identity) that finds support in philosophy, psychology, and neuroscience: the narrative identity or the narrative self. The "narrative self" is defined as a "more or less coherent self (or self-image) that is constituted with a past and a future in the various stories that we and others tell about ourselves" (see Gallagher 2000, 15). According to Paul Ricoeur (who conceptualizes this notion e.g. in 1984 and 2002), a person acts according to a personal identity that extends *and* evolves over time.[26] Ricoeur draws attention on the fundamental distinction between two uses of the concept of identity: identity as sameness (*idem*) and identity as self (*ipse*). While *idem* refers to a notion of identity of something that is always the same, immutable, permanent and unified, *ipse* refers to a notion of identity of one's selfhood through time and change. While the (inner) rational agent rests on the *idem* notion of identity, the evolving agent can be connected to both notions: I am and I am not *who* I was five years ago. Dennett (1991) proposes a version of this concept in neuroscience in which the self is defined as a "center of narrative gravity", where the various stories told about the person meet (see also Gallagher 2000). Albeit the narrative self is, in this case, seen as a fictional representation, it is an important principle of organization that through

[26] See e.g. Davis (2009) for another theory of personal identity as narrative identity. See Kirman and Teschl (2006) for a social identity perspective on the evolution of one's identity that depends on what a person currently is and does, *who* she wants to be or become, and to which social group she chooses to belong (see also Horst et al. 2006).

various narratives makes one's experience relatively coherent over extended periods of time. Changes in preferences, in this perspective, are mediated by the way one recounts to oneself the experiences that one has lived and is to live.

While the representation of an evolving agent points towards an endogenous sequence of multiple preferences, it is observationally consistent in terms of choice with an exogenous sequence of independent preferences or selves. Such representation has been sometimes used in economics. For instance, Gul and Pesendorfer (2001, 2004, 2005) model of intertemporal choice and the changing preferences model of Chapter 3 of this dissertation are consistent with the exogenous change of preferences. Similarly, Strotz (1955) interpretation of intertemporal choice illustrates an extreme version of this type of model: Instead of two preferences or selves that are in conflict at each point in time, the conflict is between today's and tomorrow's preferences. According to Strotz (1955, 179), "[t]he individual over time is an infinity of individuals".

An important feature of the evolving agent model is that it is consistent with a *person*'s reflexive capacity to evaluate and change her identity or *who* she is, wants to be or become. As argued in the previous Sections, the evolution of a person's preferences may be the result of the resolution of a conflict between first- and second-order preferences. I may, upon my experience, form a second-order preference (an identification preference) or volition against a preference that I currently hold, and in the future change my preference accordingly. But at each point in time, an individual may or may not identify with one's present preferences, as well as with the previous or next preference change. I may, for instance, not identify with a preference change or choice reversal that I feel I cannot avoid, as in the cases of relapse into addiction. The evolving agent model would accommodate, in this sense, the reflexive and non-reflexive evolution of a person's single preference.

In order to distinguish between reflexive and non-reflexive preference change, the evolution of first-order preferences must be accompanied by the evolution of the agent's second-order preferences as, for example, in the hierarchical evolving model above. In the case of models of exogenous preference change, the collection of non-choice verbal data on second-order attitudes such as identification preferences could inform if such apparently exogenous changes of preferences are the result, at least in retrospect, of reflexive or non-reflexive preference change. More structure would need to be imposed over the (hierarchical) evolving model in order to explain or predict some patterns of behavior such as the repetition of certain bias over time.

The evolving agent model seems philosophically and psychologically appealing. It may also help in centering economic analysis around meaningful questions, in particular on frequent and reasonable behavior due to preference change. The evolving agent model is not only compatible with learning and preference formation, but, also, with changing tastes or values that depend on the experience of the economic agent. Despite the appeal of the evolving agent model, one finds several economists who argue that for most purposes stable preferences are a necessary requirement for economic analysis. For instance, Samuel Bowles (1998, 79), in a survey about endogenous preferences, writes the following:

> "For preferences to have explanatory power they must be sufficiently persistent to explain behaviors over time and across situations. If preferences are endogenous with respect to economic institutions it will be important to distinguish between the effects of the incentives and constraints of an institutional setup (along with given preferences) on behaviors, and the effect of the institution on preferences per se. The key distinction is that where preferences (and not just behaviors) are endogenous they will have explanatory power in situations distinct from the institutional environments which account for their adoption. Thus, however acquired, preferences must be internalized, taking on the status of general motives or constraints on behavior. Values which become durable attributes of individuals - for example, the sense of one's own efficacy introduced below - may explain behaviors in novel situations, and hence are included in this broad concept of preferences."

The "broad concept of preferences" defended by Bowles (1998) conceptualizes preferences as attributes that are (endogenously) "acquired", but that "become permanent reasons for behavior" (Bowles 1998, 80). According to the author, only such preferences have explanatory power. Similarly, Hoff and Stiglitz (2016, 2) stress that "[p]ast social experiences and past social structures can result in sustained ways of conceptualizing a situation and, hence, sustained beliefs and social outcomes". The authors propose to focus on an "enculturated actor", whom preferences, perception, and cognition are subject to "deep" social influences rooted on the social and cultural backgrounds he is exposed to. Indeed, the dependence of preferences and behavior on "deep" social influences seems quite sensible, and a lot of empirical findings, part of them reviewed in Bowles (1998) and Hoff and Stiglitz (2016), point that this is indeed the case. I side with these authors that the endogenous determinants of preferences should be central in economics (as Chapter 5 of this dissertation testifies).

However, this is not incompatible with reflexive and non-reflexive preference change. Besides understanding the cultural and social determinants of preferences, it seems important to understand what might *change* these preferences. Market rules and economic institutions evolve over time. It is only by properly understanding why preferences change that we will be able to appraise the effect(s) of economic policies and changes in economic institutions on preferences. Carpenter (2005) provides an example. The author conducted a within-subject experiment to test if individuals' social preferences change according to different aspects of the market. His findings suggest that subjects are less pro-social in more anonymous settings due to a change in preferences. Since social preferences are likely to reflect, in some part, one's social values, it seems that these changes can be, at least in principle, the result of reflexive preference change.

In addition, it seems possible to make meaningful economic analysis even when preferences are not durable attributes of individuals. To make this point, it is useful to consider the model of changing preferences of Chapter 3. In this model, an agent *may* change preferences from one period to another, and we analyze the case in which the agent preferences may change unpredictably *at most* once. We show that an analogous condition to GARP is necessary and sufficient for the behavior to be explained by the maximization of a single preference relation between any two (not necessarily consecutive) periods, and that to rationalize one change in preferences one can apply GARP in two partitions of the time horizon. Thus, if one observes a violation of GARP between a period 1 and a given period t, the observable condition that rationalizes one change in preferences says that it is not possible to observe a second violation of GARP between t and the last period of the time horizon. Note that one could use a similar reasoning to rationalize more than one change in preferences. Then, one can assume that the behavior that satisfies GARP between any two (not necessarily consecutive) periods is the result of the maximization of a single preference. In this context, stability is not an *a-priori* assumption, but an observable and verifiable condition for any sequence of periods. In the case it is verified for any sequence of periods, it is then possible to use the consistency properties associated with stable preferences for this sequence of (time) periods. This means that the evolving agent model allows, at least in principle, to use with considerable generality and tractability many of the tools economists have developed so far.

This example illustrates that an evolving agent model can be empirically refutable. An evolving agent model would not be empirically refutable in terms of choice if no *a-priori* restriction on the number of changes of preferences is imposed, since any choice behavior whatsoever is rationalizable if we assume that the economic agent is choosing according to a single preference (possibly distinct) at every period. An evolving agent is, in this sense, trivially consistent with time- and context-dependent preferences. But as Chapter 3 of this dissertation illustrates, it is possible to impose meaningful restrictions upon choice behavior when one predicts that a limited number of changes in preferences are to occur. For instance, the changes in preferences following a change in the market rules or conditions may be anticipated and modeled as a single event of *potential* change in preferences. This means that an observer could assume, *a priori* and if desirable, that the economic agents have stable and context-independent preferences prior and after the change in the market. Brennan (1993) makes a similar point, arguing that it is worth studying preference change whenever such change is predictable:

> "Nothing in an SU [single utility] theory rules out preference changes. I will grant Professor Lutz that most economists assume preferences are stable. This assumption might be defensible because it deters us from too quickly invoking preference change to salvage a failed prediction and encourages us to base our explanations of behavior on observable phenomena such as prices or incomes. Of course, this defense can turn into demagogy. Preference change may be real and perhaps even predictable. I have elsewhere acknowledged that where preference change is likely, as in broadcasting, efficiency models may make no sense (Brennan, 1983). In such contexts, I agree with socioeconomists that economic explanations and policy evaluations should be supplemented by models of preference change." (Brennan 1993, 162)

Finally, the evolving agent model opens the possibility to search for the lower or upper bound on the number of preferences necessary to explain choice behavior. As discussed above, Kalai et al. (2002) and Apesteguia and Ballester (2010) show that this is a meaningful exercise for a model that is, in fact, observationally consistent in terms of choice with an evolving agent. Then, in cases of unpredictable number of changes in preferences, one could search for the minimal number of preference relations necessary to explain choice behavior within a given sequence of (time) periods.

7 Preferences over Preferences in Welfare Economics

> "I merely wish to emphasize here that we must look at the entire system of values, including values about values, in seeking for a truly general theory of social welfare". (Arrow 1951, 18)

One important aim of welfare economics is to provide rankings of individual and social welfare.[27] In neoclassical economics, preference satisfaction is one of the main, if not the dominant view of individual and social welfare. Even when confronted with the evidence that observed behavior differs from the maximization of a stable and context-independent preference, economists often take the satisfaction of a *given* preference relation as the benchmark for welfare analysis (e.g. Koszegi and Rabin 2007; Rubinstein and Salant 2012; Apesteguia and Ballester 2015).[28]

Apesteguia and Ballester (2015) approach to welfare illustrate this well. In their framework, an index indicates how "far" choice behavior is from the maximization of a single preference relation. Their index ranks different preference relations according to the number of alternatives in each choice problem (among the available ones) that need to be "swapped" with the chosen alternative in order to rationalize individual choices. Then, relying on standard revealed preference data, they interpret the preference ordering that "minimizes" that index (i.e., the unobservable preference relation that is "closest" to the revealed choices) as a reflection of the higher attainable individual welfare. In this sense, any decision that is inconsistent with the traditional assumptions about "rational choice" is seen to entail a welfare loss. Another example is given by Rubinstein and Salant (2012), who assume that an agent reveals distinct preference relations in different contexts that vary according to properties of the choice environment that are deemed *normatively irrelevant* (e.g. frames). They assume that the multiple preferences are the outcome of some cognitive process that distorts the agent's unobservable and underlying preference that reflects her welfare. The authors then define testable assumptions on the decision process that relate unobservable preferences to choice behavior in order to elicit the agent's underlying preference. As in the standard neoclassical approach, the satisfaction of some *context-independent* preference, in these papers *purified*[29] of mistakes, is used as a normative criterion.

Although they do not assume the existence of underlying preferences, Bernheim and Rangel (2007, 2009) build a choice-theoretic welfare criterion that relates with this view.[30] They wish to respect individuals' choices in the presence of context-dependent behavior. In order to handle such context-dependent behavior, the authors propose a *generalized choice situation* $GCS = (A, d)$ where A corresponds to a standard choice situation and d to an *ancillary condition* such as the manner in which information is presented or other frames. Like Rubinstein and Salant (2012), they deem these ancillary conditions as normatively irrelevant, i.e., "a feature of the choice environment that may affect behavior, but [that] is not taken as relevant to a social planner's evaluation" (2009, 55). They then define a welfare criterion based on what they call an "unambiguous choice relation", for which x is said to be unambiguously chosen (and welfare superior) over y if and only if y is never chosen when x is available. In this sense, only context-independent choices (that remain stable for all ancillary conditions) are considered to reveal the individuals' welfare relation. From a revealed preference perspective, this corresponds to the satisfaction of some context-independent preference relation.

The recognition that preferences change over time highlights one of the difficulties with taking the satisfaction of a given (revealed) preference relation, even when "purified" of supposed mistakes, as a measure of well-being and social welfare. Which preferences, from the several that are one own over time, should be given priority? Should or should not an observer dismiss a preference that was revealed by an agent in the past but that she no longer holds?[31] And, as Hausman (2012, 81) rightly asks, "should the consequences for welfare of a policy that changes people's preferences be measured by people's preferences before the policy is put into place or by their preferences afterward?"

Take the example of Bernheim and Rangel's (2007, 2009) welfare criterion with the ancillary condition of a time horizon $d = \{1, ..., T\}$. In their framework, if x is chosen over y at period $t-1$ and y is chosen over x at period t then x and y are said to be non-comparable in terms of welfare. In this sense, "past" $(t-1)$ and "present" (t) choices (and the potential preferences behind them) are taken *equally* in consideration.

[27] Throughout this Section I will discuss how different welfare criteria may or may not apprehend individuals' welfare, and how they *can* be useful for an observer (e.g. planner) to make inferences about individual and social welfare.

[28] See Infante et al. (2016) for a review and criticism of this approach.

[29] I borrow this term from Hausman (2012) and Infante et al. (2016).

[30] See also Fleurbaey and Schokkaert (2013) who introduce interpersonal comparisons and distributional considerations within a framework that extends Bernheim and Rangel (2007, 2009).

[31] See e.g. Parfit (1984) and Bykvist (2003) for critical analysis of these questions.

However, in many cases our intuition seems to suggest that time is normatively relevant. For instance, it seems odd to give normative authority to my childhood's preference to be a writer of poems (see Parfit 1984, ch. 8). Particularly, since today I do not identify with and/or care about being a writer of poems. More generally, it seems quite plausible to say that "[a]nyone can rationally ignore the desires that he lost because he changed his mind [desires that are no longer judged of worth or important]" (Parfit 1984, 153).

More generally, the arguments in the preceding Sections suggest that it is important to distinguish between reflexive and non-reflexive preference change. As argued before, this is possible by taking individuals' second-order preferences (when seen as identification and valuation preferences) into consideration. Whenever preferences (or choices) are not consistent with the standard neoclassical representation, identification preferences can bring information that allows one to infer if such "inconsistencies" result from reflexive or non-reflexive preference change. In the case of the childhood's preference, it seems uncontroversial that if the individual no longer identifies with such preference that this has been a reflexive preference change. In this sense, second-order preferences inform when to ignore past desires.

Valuation preferences can bring information on how a person reflexively ranks her choices in terms of value. In particular, valuation preferences can reflect *an order* of what people care about. That is, they can bring ordinal information on what an individual *values*, what is *important* to her, or what she *cares* about. This information is relevant, among other things, because "[b]esides wanting to fulfill his desire, [...] the person who cares about what he desires wants something else as well: he wants the desire to be sustained" (Frankfurt 2009, 16).

Second-order preferences may also enrich normative frameworks with forward-looking information on what people want to achieve and *who* they want to be or become. A preference (choice) that is high ranked in a person's valuation preference and that the individual does not identify with is an indication that the reversal of this preference (choice) might be an important goal for the person (who the person wants to become). And a preference (choice) that is high ranked in a person's valuation preference and that the individual identifies with is an indication that this preference (choice) might reflect who the person wants to be. As argued by Kirman and Teschl (2006, 319), "[t]he extent that a person manages to become and to be *who* she wants to be can be said to be a particular measure of her well-being and quality of life." Second-order preferences, upon some index transformation, could potentially provide such kind of "measure".

According to these arguments, second-order preferences may enrich the neoclassical approach to welfare since revealed preferences alone will often fail to reveal what constitutes welfare and what is important to people both in the present and in the future. In addition, second-order preferences (when seen as identification and valuation preferences) may have several side advantages such as "purifying" choices/preferences from the individual's *own* perspective. For instance, some identification preferences that contradict first-order preferences may refer to first-order preferences based on expensive tastes, *anti-social desires*[32], manipulation or coercion. I can have a taste for champagne, but would prefer not to prefer to have this taste. This seems to bring more confidence to an observer that, presented with the task of deciding between subsidizing two substitute goods on a limited budget, one non-expensive (say sparkling wine) and another expensive (say champagne), to subsidize the former. In this sense, the information on identification preferences would allow an observer to make an informed decision that would respect a person's evaluation about herself even if this evaluation has not caused action. Another side advantage of using second-order preferences is connected with the process of obtaining the necessary information. By asking people to state the evaluations of their own preferences (choices), the observer is letting people to think about their preferences (choices) and might led them to choose in future occasions (by their will and after a potentially slower deliberation) the option(s) that the observer *a priori* considers most favorable to them.

For example, a potential application is to use second-order preferences as a refinement to welfare rankings that combine (i) utility-based notions of well-being with (ii) survey-based data (e.g. Benjamin et al. 2014). Adding counterfactual questions regarding "what a person would prefer herself to choose" (or prefer) can inform not only on issues of self-control (as in Benjamin et al. 2012), but also on the values and goals of a person for herself (and possibly others). Questions concerning valuation preferences may bring novel information on the priorities of people over what they care and want to be or become. This would partly countervail one of the two main criticisms posed by Sen (1987, 14) against the traditional utility-based notions of welfare (either as happiness or desire-fulfillment), that "have the twin characteristics of

[32] Defined as intrinsic preferences against the well-being (or freedom) of another.

(1) being fully grounded on the mental attitude of the person, and (2) avoiding any direct reference to the person's own valuational exercise - the mental activity of valuing one kind of life rather than another."

7.1 Individual Sovereignty, Opportunities, and Context-dependency

To substantiate the usefulness of second-order preferences, I would like to discuss two criticisms of their use put forward by Robert Sugden (2004). Though they are addressed to the traditional meta-ranking view of second-order preferences, it is equally pertinent to discuss them for the present approach. The two criticisms are framed as critics to second-order preferences as an alternative approach to the author's "opportunity criterion" that is intended to respect individuals' choices without referring to the preferences that lie behind them. The first criticism is that meta-rankings based on second-order preferences are opposed to the value of opportunity and individual sovereignty. As the argument goes, "[t]he metaranking approach locates normative authority, not in the day-to-day decisions that individuals make as economic actors, but in each person's supposed higher moral self" (Sugden 2004, 1017). A robust concept of individual (consumer) sovereignty, according to this view, "should not need to invoke such a moralized account of preference". And since a second-order preference may be contrary to a first-order preference (i.e., it may value it negatively), the second-order preference may be read as a prescription for imposing restrictions on the individual's opportunities to fulfill such first-order preference.

First, I would like to argue that part of this criticism is at odds with Frankfurt's (1971, 13) view that second-order preferences/volitions are not necessarily moral:

> "In speaking of the evaluation of his own desires and motives as being characteristic of a person, I do not mean to suggest that a person's second-order volitions necessarily manifest a moral stance on his part toward his first-order desires. It may not be from the point of view of morality that the person evaluates his first-order desires. Moreover, a person may be capricious and irresponsible in forming his second-order volitions and give no serious consideration to what is at stake. Second-order volitions express evaluations only in the sense that they are preferences. There is no essential restriction on the kind of basis, if any, upon which they are formed."

According to this view, second-order preferences do not necessarily represent a "higher moral self", but instead the relative importance that a person gives to different first-order preferences that she holds. More specifically, and as argued above, second-order preferences can represent - with the exception e.g. of times when a person gives "no serious consideration to what is at stake" - what a person values and cares about. For instance, I might would like to quit smoking (even if my preference is for smoking), not because I find it the moral thing to do, but because I care about my health *or* simply because it is economically-wise. In this sense, second-order preferences are relevant not because they are moral but because they *often* reflect what a person values or cares about.

Second, I wish to argue that it is possible to take second-order preferences into consideration and still respect the value of opportunity and individual sovereignty. Take the example of Elizabeth, a smoker who has stated that would like to quit smoking (and that cares about quitting smoking). Given that preferences in general determine choices, she thus *reveals* a first-order preference for smoking, and *states* a second-order preference not to smoke. Note that this is the canonical example of a second-order preference that contradicts a first-order preference. Following Sugden (2004), an observer who wishes to respect the value of opportunity and her individual sovereignty should not prohibit smoking. This would reduce her opportunities and be against her first-order preferences. I agree.

Still, why should not Elizabeth's evaluations about her preferences be important when considering her sovereignty? Why are her revealed preferences more important than her evaluations about her preferences in this respect? Although the arguments above suggest that sometimes one should give priority to second-order preferences over first-order preferences, I wish to argue that it is possible to respect Elizabeth's opportunities and the sovereignty of her first-order preferences even when taking second-order preferences into consideration. Indeed, it is even possible to enhance Elizabeth's opportunities and her sovereignty when opportunity and sovereignty are seen from a broader "positive" perspective. To do so, it is convenient to recall the distinction between *negative* and *positive* freedoms. In its most famous formulation, due to Berlin (1969), the former is interpreted as the freedom from constraints that are imposed by others (as opposed to constraints such as economical or biological impediments).[33] For Berlin (1969, 8), negative

[33] The nature of the relevant constraints for the negative notion of freedom has been the subject of many debates on political theory. In classical debates this notion was associated with political freedom (the limits on free action that should be imposed by law) and the frontier *between the area of private life and that of public authority* (see Berlin 1969).

freedom consists in "not being prevented from choosing as I do by other men". Positive freedom, still according to Berlin (1969), consists in the ability to lead a life in an autonomous and reasoned/conscious fashion, in a way that actions are *self-directed* and not influenced by external nature and other man. In the words of Berlin (1969, 8), "[t]he freedom which consists in being one's own master". In economics, Sen (1988) favors a notion of positive freedom that consists in what a person is able to choose to do or achieve. This corresponds to focus on the actual or "real" opportunity to choose, rather than on the absence of constraints to achieve certain goals.

Now, if we interpret respecting Elizabeth's opportunities and sovereignty in *positive* terms (i.e., including both negative and positive freedoms), it is possible to take second-order preferences into account and respect or even enhance, in this sense, her opportunities and the sovereignty of her preferences. In particular, if an observer wishes to respect the sovereignty of Elizabeth's first- and second-order preferences they should not only *not* prohibit smoking *but also* enhance some policy that would help her to surpass what *she* considers to be a negative preference/behavior (or her "weakness of will"/addiction, since her behavior is not consistent with her will). For instance, providing Elizabeth with free consultation(s) with a specialized doctor could be such a policy. By doing so, the observer would still respect the value of opportunity (in terms of negative freedom), while in addition enhancing her positive freedom to lead the life that according to *her* second-order preferences she has a reason to value. Without such policy, Elizabeth has the negative freedom to go to a consultation (it suffices that it is available on the market), but she may not have the positive freedom (or "real" opportunity/"capability") to do so, for, say, budget constraints. Since the two policies (non-prohibition and free consultation) are not mutually exclusive, the observer would *respect* Elizabeth's sovereignty in terms of her first- and second-order preferences, which seems to favor a broader/positive view over individual sovereignty.

The second criticism of second-order preferences (but also of preference satisfaction in general) presented by Sugden (2004) is based on the evidence that preferences are often susceptible of changing according to *trivial changes in viewpoint or context*. As the argument goes, as soon as we acknowledge that preferences are unstable or context-dependent, using preference satisfaction (either of first- or second-order) for normative analysis is not possible:

> "Economists often want to make normative comparisons between very different social states - for example, between a future in which international trade is subject to tariffs and one in which it is not. The standard methods of welfare economics hold individuals' preferences constant across the relevant social states, treat those constant preferences as measures of well-being, and ask how far they are satisfied in each state. Such analysis is not possible if individuals' preferences shift according to trivial changes in viewpoint or context." (Sugden 2004, 1016)

Sugden (2004) is right in arguing that if preferences are totally contingent upon *trivial changes in viewpoint or context*, then normative analysis is better done without taking preferences into consideration. Still, there seem to remain cases in which preferences change for predictable and reasonable reasons. As argued in Section 6.2, preferences may change after some predictable or known event; they can also change because of the resolution of a conflict between first- and second-order preferences (e.g. Elizabeth adopts a first-order preference not to smoke, and consequently stops smoking).

In addition, as the literature on the measurement of freedom and the ranking of opportunity sets illustrates (e.g. Pattanaik and Xu 1990; Foster 1993; Gravel 1994; Nehring and Puppe 1999)[34], excluding information on individual preferences may leave us with coarse criteria. In the setting of ranking opportunity sets according to the freedom they offer, such an approach often leads to rankings based on the number of alternatives of each set (e.g. Pattanaik and Xu 1990). However, it is questionable that a set containing two "good" alternatives provides the same amount of freedom than a set containing two "bad" alternatives.[35]

Despite these arguments in favor of using preference information, I side with Sugden (2004) in that objective measures of well-being (such as opportunities) are necessary and should be central in welfare analysis (see also Section 8.4 below). Preferences (either of first- or second-order) are based on what people want, their desires and goals. It follows that alternatives that are not the focus of the desires, wants, or

[34] An opportunity set is any set of alternatives (assumed to be mutually exclusive) that are available for choice for an individual. The main question in this setting is what it means - according to a definition of freedom or opportunity - for one opportunity set to offer more freedom/opportunity than another. See Gravel (2008) for a comprehensive review of the use of the notion of freedom in economics.

[35] This discussion is often centered around one of the axioms of Pattanaik and Xu's (1990) ranking of opportunity sets that states that all opportunity sets that contain one alternative (singletons) offer the same amount of freedom of choice. See Jones and Sugden (1982) and Sen (1991) for competing views.

goals of individuals have no normative authority. Then, as pointed by Sen (1980, 210), "[o]pportunities have no value in a desire-supported system, only *desires* for opportunities have, and objective contraction of opportunities can be washed out by subjective change of desires." Further, I believe that Sugden's (2004) arguments bring more strength to this view. Given the potential contingency of preferences (either of first- or second-order) on trivial changes in viewpoint or context, it will sometimes be difficult to have credible rankings based on preference information. Nonetheless, what I intend to suggest in this Section is that we may not want to forget about the information on preferences all together, but try instead to have richer data sets that include information on first- and second-order preferences. I turn now to a potential application that would use a data set enriched in this sense.

7.2 An Application to Bernheim and Rangel (2007, 2009)

Information on second-order preferences can be used as what Bernheim and Rangel (2007, 2009) call a *refinement* of their welfare ranking. As pointed by Rubinstein and Salant (2012) and others, Bernheim and Rangel (2007, 2009) welfare ranking is typically a coarse binary relation that becomes more so as the number of choice observations increases. For that reason, Bernheim and Rangel (2007, 2009) propose to use "nonchoice evidence [such as evidence on inattention] to officiate between conflicting choice data by deleting suspect GCSs" (2007, 469). In this Section, I use the retrospective model and identification preferences to select "suspect" $GCSs = (A, d)$.[36] The feature of the choice environment d is then the periods at which the choices have been made. In the framework of the retrospective model, Bernheim and Rangel's (2009) preferred welfare criterion can be written as follows:

Definition 8 Alternative x is said to be **welfare choice-superior** to alternative y if and only if $y \neq C(t, A(t))$ for all $A(t) \in X$ such that $x, y \in A(t)$.

Identification preferences provide a non-arbitrary justification to exclude GCSs based on the individuals' *own* evaluations/judgments of their choices. It is also reasonable to take them into account if one wishes to respect individual sovereignty in the broad (positive) sense defended here. For example, if when provoked I harmed another man when I would have preferred to stay calm, it is suspect to consider that the choice of harming the other man (as opposed to staying calm) was the best in terms of my well-being. Similarly, if between two cellphone alternatives I buy an expensive one (as opposed to an economical one) but asked about this choice I state that it was a mistake (because, say, I value and identify myself with being frugal but was unable to be so because of the beautiful commercial add in favor of the expensive cellphone), it seems at least prudent not to consider this choice to reflect what is best for my well-being.

In order to respect individual sovereignty one would then delete (or not) GCSs based on the agent's identification preferences. As an illustration, take the example of binary choices. One can then distinguish two cases. The first is when an individual does not identify with a given choice (i.e., the choice is a self-authenticated mistake). In this case, even if x is *consistently chosen* over y there is an argument to not count x as being welfare superior to y. Notice that this is a prudent refinement that would turn a welfare ranking based on GCSs *coarser*[37]. The second is when an individual is inconsistent in her choices but identifies with one choice (say of x over y) and not with the other (of y over x). In this case, even if y has been chosen over x, there is an argument to count x as being welfare superior to y. Notice that in this case this refinement would turn the welfare ranking *finer*. An example of the second sort of cases is when an agent identifies with preference change. For instance, an agent loves meat but becomes a vegetarian for ethical reasons. At period T, she does not identify with her past choices of meat. Then, there is an argument to only take the later choice for vegetables into account even if meat has been chosen in the past.

Since we are looking at choices over subsets of any size, it is possible to have many different welfare criteria that take identification preferences into account. I give two examples based on the definitions introduced in Section 5: One that respects reflexive choices, and another that discards mistakes. The former can be stated as follows:

[36] For the sake of presentation I abstract from valuation preferences, but they could bring meaningful information, for example, in terms of priorities over policy measures associated with this kind of welfare ranking.

[37] A ranking becomes coarser (finer) when it becomes less (more) discerning. Formally $\succsim$ is said to be coarser than $\succsim'$ if $x \succsim' y$ implies $x \succsim y$, and if $\succsim$ is coarser than $\succsim'$ then $\succsim'$ is said to be finer than $\succsim$ (see e.g. Bernheim and Rangel 2007, 469).

Definition 9 Alternative x is said to be **welfare reflexive-superior** to alternative y if and only if (i) $x = C(t, A(t))$ for some $A(t) \in X$ such that $x, y \in A(t)$, (ii) $x \succsim_t^I w$ for some distinct $w \in A(t)$ and for all $A(t) \in X$ such that $x, y \in A(t)$ and $x = C(t, A(t))$, and (iii) there exists some distinct $z \in A(t)$ such that $z \succ_t^I y$ and/or there exist no $v \in A(t)$ such that $y \succsim_t^I v$ for all $A(t) \in X$ such that $x, y \in A(t)$ and $y = C(t, A(t))$.

In words, x is said to be (welfare) reflexive-superior to alternative y if x is chosen at least once when y is present, if whenever x is chosen and y is present the choice is reflexive, and if whenever y is chosen and x is present the choice is *not* reflexive. Remark that this definition does not entail that for x to be welfare superior to y that x is *reflexive-consistently* chosen over y. It may be the case that y has been chosen for some $A(t) \in X$ such that $x, y \in A(t)$. The condition is that in that case the choice is not reflexive (but not necessarily a self-authenticated mistake). Remark also that in case of indifference in terms of identification over one's choice the criterion does not discard this choice. Take the example of binary choices. If y is chosen once over x and is indifferent in terms of identification to x in that period and x is chosen once over y and the agent identifies with that choice, then the ranking does not compare both alternatives in terms of welfare. Finally, note that this criterion, contrary to the one by Bernheim and Rangel's (2007, 2009), requires that $x = C(A(t))$ for some $A(t) \in X$ such that $x, y \in A(t)$ in order for x to be considered welfare superior to y. The underlying reason for this difference is that Bernheim and Rangel's (2007, 2009) preferred welfare criterion depends on defining the choice domain to include every non-empty finite subset of X. In fact, if we do not observe an agent's choices from all these subsets (or at least from all pairs), as it will be most often the case, their preferred welfare choice-relation can be said to be less appealing. It could be the case that x would be said to be *welfare choice-indifferent* to y (as opposed to incomparable) even though x and y have never been compared in terms of choice.

I now state a criterion that instead of respecting reflexive choices discards mistakes:

Definition 10 Alternative x is said to be **welfare self-superior** to alternative y if and only if (i) $x = C(t, A(t))$ for some $A(t) \in X$ such that $x, y \in A(t)$, (ii) there exist no distinct $w \in A(t)$ such that $w \succ_t^I x$ for all $A(t) \in X$ such that $x, y \in A(t)$ and $x = C(t, A(t))$, and (iii) there exist some distinct $z \in A(t)$ such that $z \succ_t^I y$ for all $A(t) \in X$ such that $x, y \in A(t)$ and $y = C(t, A(t))$.

In words, x is said to be (welfare) self-superior to alternative y if x is chosen at least once when y is present, if whenever x is chosen and y is present the choice is not a self-authenticated mistake, and if whenever y is chosen and x is present the choice is a mistake. In the case of binary choices over x and y, this means that whenever y is chosen the agent identifies more with the choice of x. Remark, however, that the same is not necessarily the case for choices from subsets with more than two alternatives.

Under this general domain, these two welfare rankings (formulated in Definitions 9-10) are not necessarily complete, and more importantly, not necessarily acyclic. This may be problematic since without acyclicity it *may* not be possible to identify maximal alternatives for finite sets and/or unambiguous welfare improvements. Bernheim and Rangel's (2007, 2009) preferred welfare criterion (analogous to Definition 8) is also cyclic under this general domain. Their preferred welfare criterion is only acyclic when considering a choice domain that includes all conceivable choice problems of X. An open question remains concerning which restrictions upon the binary relations and/or the choice domain would need to be imposed in order for the reflexive and self welfare rankings to be acyclic.

Finally, remark that identification preferences may be inconsistent over the same choice across time (e.g. $y \succ_t^I x$ and $x \succ_T^I y$). At T, I may say that when I was a child I identified with my desire to become a poet instead of a researcher (say $y \succ_t^I x$), while today I do not identify with such preference (say $x \succ_T^I y$). This unveils one potential issue with posing question concerning identification or of the type "what would you want yourself to have chosen?" when the time dimension is relevant. In this example, it seems that I have changed my values/second-order preferences, but it is possible that I would not want myself to have chosen differently in the past. In this kind of cases, the reflexive-choice criterion would deem the two alternatives as non-comparable in terms of welfare, when it seems that at least from the present perspective x is welfare superior to y.

One way to deal with these cases would be to, whenever identification preferences are not consistent over time, give priority to identification preferences over later (as opposed to former) choices. This could have the additional advantage of turning the welfare ranking finer. For instance, suppose that a person states that she identifies with the choice of x over y at period t, but she states that she does not identify with the choice of x over y from period $t + 1$ onward. This seems to represent the childhood/adulthood preferences mentioned earlier. In this case, an observer interested in more discretion could take the later identification preferences as reflecting the (present) values of the person.

7.3 An Application to Intertemporal Preference Reversals

As a final illustration of the normative implications of reflexive preferences, consider again the case of preference reversals in intertemporal choice between a smaller short-term reward and a larger long-term reward. Consider the following example, taken from Gul and Pesendorfer (2008, 30-2). There are three consumption periods ($t = 1, 2, 3$), and three possible consumption paths (c_1, c_2, c_3): $(0, 0, 9)$, $(1, 0, 0)$, and $(0, 3, 0)$. In period 1, the agent chooses [prefers] $(0, 0, 9)$ over $(1, 0, 0)$ and $(1, 0, 0)$ over $(0, 3, 0)$. And in period 2, the agent chooses $(0, 3, 0)$ over $(0, 0, 9)$. Now suppose the agent faces the following decision problem: she can either choose $(1, 0, 0)$ in period 1 or leave the choice between $(0, 0, 9)$ and $(0, 3, 0)$ for period 2. *According to her preferences at period 1*, she chooses $(1, 0, 0)$. In fact, if she does not "commit" to this choice at period 1 and leaves the choice to period 2 between $(0, 0, 9)$ and $(0, 3, 0)$, she will end up choosing $(0, 3, 0)$ at period 2 which, although it is her most preferred option at period 2, it is the less preferred option from the point of view of period 1.

Gul and Pesendorfer (2008, 30), based on their previous work in Gul and Pesendorfer (2001, 2004, 2005), endorse a "standard, single-self model that accounts for this behavior". Denote by $\mathcal{C}$ the set of second-period choice problems, where $C \in \mathcal{C}$ consists of a consumption path with identical first-period consumption. Choosing $(1, 0, 0)$ in period 1 corresponds to $\{(1, 0, 0)\}$, while leaving the choice for period 2 consists of $C = \{(0, 3, 0), (0, 0, 9)\}$. Then, the authors describe period 1 preferences as follows:

$$\{(0, 0, 9)\} \succ_1 \{(1, 0, 0)\} \succ_1 C = \{(0, 3, 0), (0, 0, 9)\} \sim_1 \{(0, 3, 0)\}$$

where $\succ_1$ denotes a strict preference and $\sim_1$ indifference in period 1. This is consistent with the above preferences/behavior, since $\{(1, 0, 0)\}$ is ranked above the second period choice of $C = \{(0, 3, 0), (0, 0, 9)\}$. According to the authors, "[p]eriod 1 behavior reveals that the individual's welfare is higher in all periods when she is committed to $(0, 0, 9)$ than when she must choose from C in period 2." (Gul and Pesendorfer 2008, 31).

The authors contrast this representation with the two-parameter model that modifies exponential discounting reviewed in Section 3 (e.g. Laibson 1997; O'Donoghue and Rabin 1999, 2001, 2003). Taking the preceding three-period decision problem, the agent's instantaneous utility for each period, u_t, can be represented as follows:

$$\begin{aligned}
u_1(c_1, c_2, c_3) &= c_1 + \beta\delta c_2 + \beta\delta^2 c_3 \\
u_2(c_1, c_2, c_3) &= c_2 + \beta\delta c_3 \\
u_3(c_1, c_2, c_3) &= \delta c_3
\end{aligned}$$

where $\beta > 0$ and $\delta < 1$. Then, according to the discussion above on the normative authority of an *inner or outer rational agent*, it is common practice to take the long-run perspective as the right welfare criterion (as e.g. in O'Donoghue and Rabin 1999, 2003). This corresponds to set $\beta = 1$, which yields the following *fictitious* utility function:

$$u_0(c_1, c_2, c_3) = c_1 + \delta c_2 + \delta^2 c_3$$

which is interpreted as the agent's reconstructed preferences would she not been distorted by a faulty psychologically bias towards the present (e.g. O'Donoghue and Rabin 2003). Gul and Pesendorfer (2008, 31) argue that this welfare criterion is quite arbitrary if one interprets each utility function as a different "self", since, for example, it "assigns a higher welfare to $(1, 0, 11)$ than to $(2, 3, 0)$ even though selves 1 and 2 prefer $(2, 3, 0)$". In my view, this welfare criterion seems somewhat arbitrary because it is not clear why β should be exactly equal to 1 even if one interprets the present bias as a defect.[38] But most importantly, it seems that this welfare criterion applies a *one-size-fits-all* solution that may not be adapted for every person. As argued above, while the behavior that a present bias entails may be indeed considered as a mistake by some persons, it may *not* be by others. This, as argued above, can be conceptualized through the distinction between reflexive and non-reflexive preference change.

There are two possible cases (considering stable identification preferences). The person does not identify with the present bias, and herself does not want it to be her will. She, and not the observer,

[38] O'Donoghue and Rabin (1999, 113-4) argue that from a long-run perspective, even if β is very close to 1 it can create "an arbitrarily large welfare loss" since the rewards and costs can be arbitrarily large or because it is possible to have finitely many periods. It seems to me that this is an insufficient justification for taking $\beta = 1$ as the appropriate welfare criterion for *all* cases (e.g. in a case with few periods and small costs and rewards).

considers it to be a defect. In this case it seems sensible to consider a welfare criterion where β is equal or close to 1. If one is sensible to Sugden's (2004) arguments in favor of the sovereignty of first-order preferences and the value of opportunity, one may not endorse any policy that restricts the person's opportunities. However, one may still favor a policy that helps the person to overcome what *she* considers to be a defect. The other possibility is that the agent identifies with the present bias. In this case, the choice to "overrule" a person's preferences (and her choices) is even more controversial than in the standard case without information on second-order preferences. Having the individuals' evaluation of their preferences may then help us to satisfy their sovereignty and the value of opportunity. In addition, either by helping individuals to overcome what they consider to be a defect or abstaining from doing so in cases they do not consider it to be a defect, second-order preferences may provide relevant information to direct resources to people's goals, what they value and seem to care about.

This means that, at least in principle, it is possible to use a multiple preferences approach without constructing a *paternalistic welfare criterion*. By recognizing the person's evolution over time, and the relevance of her evaluation of her preferences, one may respect not only the sovereignty of her preferences but also the sovereignty of her evaluation of those preferences.

8 Discussion

In this Section I discuss some features, limitations, and potential extensions of the use of second-order preferences in economics. I start with some comments on how to recover second-order preferences (Section 8.1), then I briefly discuss the notion of identification (Section 8.2), and continue with some of the potential limitations of this framework due to adaptation and false beliefs (Section 8.3). I then discuss some limitations and potential extensions for welfare economics (Section 8.4), and finish by distinguishing the view defended here from the inner rational view of agency (Section 8.5).

8.1 Data

An important feature of any agential theory in economics is how to recover the objects of interest. Traditionally, economics has used choice (or stated choice) as the main source of data. This has led to the identification of the behavioral implications of many theories of choice (often in the form of revealed preference-like axioms). Though in some occasions second-order preferences will translate into choice behavior, other times they will not be reflected on the agents' choices. In this Section, I briefly discuss a survey-based and a choice-based method to recover second-order preferences.

Survey-based data, such as individuals' verbal evaluations of their preferences or choices, seems to be one of the most immediate ways of how to recover (evaluative) judgments of an individual. However, as with first-order preferences and choices, second-order preferences may be prone to be affected by frames, cognitive bias, and other sources of context-dependency that limit the reliability of the data collected. If more or less than first-order preferences is an empirical question. But there is sufficient evidence that judgments are prone to bias.[39] For example, Schkade and Kahneman (1998) observed that while students from two Midwest and two California Universities believed that students in California would be significantly happier, the self-reported happiness was very similar in the two locations. This example illustrates a bias/misprediction in judgments concerning adaptation to ways or places of living. They explain this bias through a *focusing illusion*: when reporting their well-being students focused on central aspects of life, while when imagining the happiness of someone else in a different location they focused on the dimensions that differ across regions (in this case, climate). They conclude that "[n]othing in life matters quite as much as you think it does while you are thinking about it".

Similar concerns apply to evaluations of past experiences. There is by now some evidence that evaluations of remembered hedonic utility are anchored on the individuals' emotional state when the evaluation takes place (see Stone and Shiffman 1994). For example, high levels of a measure (e.g. pain) on the day of the retrospective evaluation may upwardly bias the retrospective recall of that measure made on that day (Stone et al. 1997, 186). If these bias extend to the evaluation of past choices, desires, or preferences is again an empirical question. But this evidence suggests that care should be taken in the design and interpretation of applications of the hierarchical retrospective model.

[39] See Kahneman and Thaler (2006) for a review of empirical findings on bias on forecasting/remembering experienced (hedonic) utility.

In addition, survey-based data is often non-incentivized, which may favor inattention and deception. An interesting and controlled way to circumvent some of these limitations is through survey-based experiments. For example, in an experimental setting it is possible to record (in a systematic and controlled way) the duration taken to give an answer in order to exclude *speedy* answers that cannot be the result of honest attentive answers. Another example is the inclusion of incentivized questions of comprehension, which may favor subjects' attention. Though this type of procedures are not perfect solutions, they may increase (if associated with other measures) the reliability of non-choice data.

In a recent survey-based experiment, that brings some evidence that welfare rankings based on (stated) choices are consistent with the happiness view of utility, Benjamin et al. (2012) asked subjects their meta-choices (identification preferences) over binary stated choices in a series of hypothetical choice scenarios. For example, one of their hypothetical choice scenarios was between a job in which the subject would "sleep more but earn less" and a job in which the subject would "sleep less but earn more". Then, for each scenario, they asked subjects the two following questions: "If you were limited to these two options, which do you think you would choose?" (stated choice), followed by "If you were limited to these two options, which would you want yourself to choose?" (stated meta-choice). Interestingly enough, and bringing some evidence that the conflict between first- and second-order preferences is meaningful, 28% of subjects' stated choices conflicted with their stated meta-choices.[40]

In this experiment there was no time horizon as in the hierarchical preferences models of Section 5. But in principle, the elicitation of second-order preferences (either of identification or valuation) can be done either *ex-ante* or in retrospect for most preferences and choices. Note that hypothetical stated choices may sometimes elicit meta-choices, i.e., they may elicit not what the agent's would actually choose when presented with the choice situation but what they would like themselves to choose. Though separating the two questions as in Benjamin et al. (2012) may cue subjects to distinguish between these two notions and give a reasoned answer, this is a problem to have in mind in a survey-based method that pertains to elicit first- and second-order attitudes.

An alternative choice-based method to elicit second-order preferences is with data on *precommitment*, defined as the deliberate restriction of a feasible set (Elster 1982, 222). George (1984, 97) labels these actions as "self-paternalistic", in the sense of an action that "an agent undertakes with the intent of reducing in some way the choice set that he will face at some future time". As the author notes, "the 'revelation' of a meta-preference may be understood to occur via acts of self-paternalism" (p. 95). Though this is certainly not the only behavioral implication of second-order preferences, it seems an important case that can be explained by the conflict between first- and second-order preferences.[41] Choice of precommitment may also be interpreted as an indication that a second-order preference (e.g. in favor of abstaining from smoking) has more value than a first-order preference (e.g. in favor of smoking) (see also Jeffrey 1974, 383). Combining data on precommitment and retrospective evaluations of choices may be an interesting avenue of research. This could, among other things, bring some evidence on how second-order preferences and self-paternalism are related with regret.

8.2 Identification

Throughout this Chapter I have used the notion of identification quite broadly, but its precise meaning is somewhat more elusive. Does it mean (precisely) that a person judges a preference/choice as if it is her own in some strong sense? Does it mean that a person judges a preference/choice to reflect who she is or wants to be or become? Or just that a person judges (or even feels) a preference/choice not to be external, refusing sentences of the type "I was not the one to do x". I do not wish to provide here a precise definition, but to briefly revise some of the notions proposed in the literature that are related with second-order attitudes. Among other things, this may help the design of questions for survey-based experiments in order to elicit identification with observed choices.

[40] In their payed and more controlled experiment run with Cornell University students the percentage increases to 33%. In that experiment some examples of reasons given by subjects for this conflict are: "Sometimes what I want to do may not be what should I do. The choice I should make may be financially better, for instance, but may not be the one I want to choose."; "I made choices I would ideally not want myself to choose in order to secure my future or make friends/family happier. I would probably regret these choices for the reason that life's too short."; "Based on long-term, overall benefits."; "Sometimes I wish my priorities were different than they actually are.".

[41] See George (1984, 96-100) for why other explanations of precommitment, namely (i) that an agent believes he would be "unable" to choose what he *prefers* (instead of what she would want herself to prefer) and (ii) the conflict between an impulsive and a far-sighted self, create difficulties in terms of choice determination and welfare analysis respectively.

Though in his early work Frankfurt (1971) seemed to lump together second-order volitions and identification, there is now some agreement that identification consists of higher-order desires/volitions and something more (see Bratman (1996) and Fischer (1999) for reviews). One of the reasons for this is that, as argued above, it is possible that a person does not care and gives no serious consideration to what is at stake when forming a given second-order volition/preference. Another is that one can always refer to a higher-order volition/preference to question if a lower-order volition/preference is really one's own in the strong sense identification seems to require (see e.g. Bratman 2003).[42] Frankfurt himself, in later articles, proposed two separate possibilities for the feature that should endorse higher-order attitudes in order to capture identification (and counter these criticisms).

The first was to say that in order to identify with a desire (choice) the second-order attitudes towards this desire (choice) needed to be *wholehearted*. According to Frankfurt (1988), a person acts wholeheartedly when she has "made up her mind" about a decision to take. This requires (roughly) that the decision is *decisive*, in the sense that the person judges that no further consultation of higher-order preferences is needed (see also Frankfurt 1971, 16). The will is undivided and the person is volitionally unified (see also Frankfurt 2009, 91-5). If without reservation or conflict I value and decided to follow my preference to be faithful to my wife, according to this view it seems that no question remains concerning if I value to value (a third-order preference) to be faithful.

Later, the author proposed another criterion based on the *satisfaction* with higher-order attitudes. In his own words, "identification is constituted neatly by an endorsing higher-order desire with which the person is satisfied" (Frankfurt 1992, 14). This requires (again roughly) that a person is settled with respect to the higher-order attitudes over a desire (choice), in the sense that the person *has no interest* in making changes to these higher-order attitudes. It differs from being wholehearted in the sense that it is not an active decision, but just a state in which questions concerning the desirability of the relevant higher-order attitudes do not arise.

An alternative criterion, proposed by Bratman (e.g. 1996), is to consider an endorsement of second-order attitudes based on a decision to treat a desire "as reason-giving in one's practical reasoning and planning concerning some relevant circumstances" (p. 9). According to the author, "[t]o identify with one's desire is (a) to reach a decision to treat that desire as reason-giving and to be satisfied with that decision, and (b_1) to treat that desire as reason-giving or, at least, (b_2) to be fully prepared to treat it as reason-giving were a relevant occasion to arise" (Bratman 1996, 12). Though all these criteria seem to have their own shortcomings, the discussion around identification suggests that the following is true:

> "It appears, then, that the special status of higher-order volitions must be explained by something other than the fact that they are desires of a higher order. It must be explained by the fact that they are endorsed by the agent in acts of identification or decisive commitments." (Lippert-Rasmussen 2003, 354; see also Watson 1975, 217-9)

8.3 Adaptive and Informational Preference Change

It is important to distinguish preference change due to the resolution of a conflict between a first- and second-order preference (reflexive preference change), from other phenomenons that also tend to change preferences over time. I will briefly discuss two of these phenomenons here: (i) adaptive preference change and (ii) informational preference change. In what follows, I draw upon Elster (1982) that is a classical (and in many ways comprehensive) treatment of adaptive preferences, and Cowen (1993) and Harsanyi (1997) for treatments of the question of preference change due to learning and/or experience.

Adaptive preference change, according to Elster (1982), refers to the cases in which aspirations are downgraded due to restrictions in the feasible set.[43] For example, say someone prefers job x that she can get if promoted to her current job y. Before the decision of promotion takes place, her feasible set of possible outcomes (at least in terms of her aspirations) is $\{x, y\}$. But if she does not get the promotion, her feasible set (in this sense) becomes $\{y\}$. Then, if that is the case, she may rationalize the non-promotion

[42] Clearly every choice is one's own in an important (even if) trivial sense. See Watson (1975, 217-9) for the origin of these criticisms.

[43] This contrasts with a preference for what one does not have (e.g. I prefer to be single to be married when I am married, but I prefer to be married to be single when I am single). Elster (1982, 226) also distinguishes adaptive preference change "from learning in that it is reversible; from precommitment in that it is an effect and not a cause of a restricted feasible set; from manipulation in that it is endogenous; from character planning [reflexive preference change] in that it is causal; and from wishful thinking in that it concerns the evaluation rather than the perception of the situation".

by saying that "the top job is not worth having anyway", changing her preferences for y over x (see Elster 1982, 225).

As exposed by Elster (1982), adaptive preferences (or preference change) differs in important ways from reflexive preferences (or change). The first distinction is that adaptive preference formation or change takes place "behind the back" of the person, as a causal (non-conscious) drive, while reflexive preference change (or deliberate character planning in Elster 1982, 224) is in general deliberate, conscious, and intentional.

Another distinction, according to Elster (1982, 237-8), is that in order to treat adaptive preferences one needs to consider the "*genesis* of wants", which involves "an inquiry into the history of the actual preferences". This means looking at the sequence of choices as revealing information about the *nature* of these choices. Here, I have taken the nature of choices to be mostly revealed through stated second-order attitudes (at least when one is interested in distinguishing between reflexive and non-reflexive choices). One could instead use the sequence of first- and second-order preferences in order to try to capture some patterns that give us indications if preferences have been reflexive, adaptive, or other. The potential of such line of inquiry is mostly an open question.

A third and important distinction, according to Elster (1982, 235), is that while reflexive preference change may "improve welfare without loss of autonomy", that is not the case of adaptive preference change. Even though adaptive preferences may improve welfare (e.g. due to resignation and reduction of frustration), they do so in a non-autonomous way.[44] Elster (1982, 235) recognizes that second-order attitudes may not be autonomous, but argues that such cases are not centrally important. I think that such cases should be taken seriously, since, as argued in Section 8.1, evaluative judgments are also prone to adaptation and other bias.

Informational preference change refers to the cases in which preferences change due to learning and/or experience. For example, a patient may prefer treatment x over y based on false beliefs about the secondary effects of treatment x. Informed about these effects, she may change her preferences for y over x. This is different from reflexive preference change since it does not *necessarily* relate with identification, nor identification necessarily relates with new information, though new information may be one way in which one changes her identification over some preferences/choices. This kind of examples led many authors to endorse different versions of fully-informed preferences for policy and welfare evaluations.[45] It is then useful to distinguish fully-informed preferences from reflexive preferences.

There are at least two ways of defining fully-informed preferences. I call them the *hypothetical informed preferences* and *actual informed preferences*. The former are the *hypothetical* preferences the agent would have if she had all the relevant information and made full use of this information (e.g. Harsanyi 1997, 133). Sometimes this version of informed preferences is coupled with some demands of rationality in terms of the use of information. The latter are the individual preferences the agent would *actually* have if she would be informed of all the relevant information. This version therefore involves no demand of full or rational use of information. Actual informed preferences are the agent's own preferences after being informed of all the relevant information. This relevant information may include info on cognitive bias that often affect agents' preferences, such as framing or adaptation.

Informed preferences face some difficulties of their own. For one, it is not necessarily straightforward to define what is the *relevant* information. How much and what type of information is necessary for one to be fully-informed about some topic? Second, informed preferences, either actual or hypothetical, are very often not available. They do not exist *now* as they pertain to the preferences with information (and cognitive capacities in the hypothetical case) that the agent does not have:

> "The preferences of perfectly informed individuals are not always relevant for imperfectly informed choice. By considering perfectly informed preferences, we are hypothetically changing an individual's human capital endowment. What an individual would want with a different human capital endowment cannot necessarily be extrapolated usefully into information about what improves the welfare of an individual now" (Cowen 1993, 262)

Finally, informed preferences, in particular hypothetical ones, may be against individual sovereignty. Contrary to reflexive preferences, informed preferences refer to preferences that are not of the individuals themselves. For example, Harsanyi (1997) defines "mistaken preferences" as actual preferences that are based on some objective view of incorrect or incomplete information. Then, one of his proposals is to

[44] Though Elster (1982) does not provide a definition or criterion for autonomous wants, he consider adaptation (in the sense defined above) as a mechanism that shapes individual wants in a non-autonomous way (see p. 228).

[45] From which Hausman (2012) is a notable example. See Cowen (1993) for an old but interesting review.

base the welfare evaluation of one individual on the "preferences of *other* knowledgeable people" (see Harsanyi 1997, 134 and 142-3). But one can certainly find many examples where the preferences of most knowledgeable people do not respect one's preferences.[46]

Despite these limitations, informed preferences have the potential to "purify" preferences of cases of blatant false beliefs, as the example of medical treatments highlights. This is something that, as it was the case with adaptation, escapes from the notion of reflexive preferences. Then, a full-fledged theory of preferences and well-being may need some kind of adaptation and informational criteria if they are to be used for policy and welfare analysis.

One standard strategy, as suggested by Harsanyi (1997), is to define some actions as *objective mistakes*, i.e., to judge some actions as against the agent's own interest even if the agent does not necessarily agree. A similar strategy is to use an observer's meta-ranking. According to Amartya Sen (1974, 1977, 1980), instead of simply looking to the persons' multiple preferences over alternatives we should also rank their multiple preferences according to some social desirable criteria. If preferences based on adaptation and false beliefs are judged negatively for a given context, they could be low ranked in a partial or complete ordering of different preferences according to their social or moral worth. According to Sen (1977, 338), this kind of meta-rankings would endow an observer with "a varying extent of moral articulation". An observer's ranking over the multiple preferences could, if information and confidence on social judgments are sufficient, rank in a social or moral convenient way preferences based on false beliefs and adaptation, as well as preferences based on expensive tastes, antisocial desires, coercion or manipulation.

But these strategies are at the cost of individual sovereignty. I have tried to design a model where mistakes are instead self-authenticated. This could, at least in principle, be endorsed both by author in favor or against paternalism. I have also defended that judgments, much like choices, are important if we want to respect a broad (positive) view of individual sovereignty. So how to to save this model in face of uninformed and adaptive preferences?

One circuitous strategy would be to define the adaptation and informational criteria with respect to the *context* of interest, i.e., to define a minimum degree of information and non-adaptation in order to use first- and second-order preferences as guidance for positive and normative analysis. This would allow to keep an hierarchical structure without defining objective mistakes with respect to observed choices as long as the context is "information and adaptation proof". But of course, this is silent with respect to the other contexts where lack of crucial information and adaptation seem to be important.

Another possibility for those, as myself, that care about individual sovereignty but think that false beliefs and adaptation are important and real issues would be to respect first- and second-order preferences at the same time as implementing "experimentation" (non-coercive and non-obligatory) policies aimed to deal with these issues. Elster (1982, 221) gives an example to deal with lack of information with respect to different ways of living: "a systematic policy of experimentation that gave individuals an opportunity to learn about new alternatives without definite commitment". Of course, in an economy with limited resources, it may be that this kind of policies may be only implemented at the expense of some others that favor the actual preferences and judgments of individuals (implying a kind of negative paternalism; see footnote 46). This kind of policy could also procure a loss in terms of the surprise and discovery that certain goods procure to an individual (see Cowen 1993, 263). What to do is a hard question that I wish to leave open.

8.4 The Objective and Intrinsic Values of Opportunity

As suggested in the previous Sections, the satisfaction of preferences may not be a good indicator of individual and social welfare for a number of reasons besides the conflict between first- and second-order preferences. Adaptation, false beliefs, or antisocial preferences are just a few examples of the difficulties with this stand (see e.g. Hausman 2012, 81-2). In addition, two important dimensions that a normative framework based on information on first- and second-order preferences would fall short are: (i) the *objective value* of opportunity, and (ii) the *intrinsic value* of opportunity. With respect to the first, by relying exclusively on first- and second-order preferences it is possible to have welfare criteria compatible with deficient opportunities, individual rights, or basic capabilities. This is an uncontroversial deficiency of a normative framework for the ones, as myself, that believe that welfare criteria based on preferences

[46] Harsanyi (1997, see 134) defends a *negative* version of paternalism, in which *we* do not coerce another into what we think is the correct behavior but we only refuse to subsidize activities that we consider to be against the agent's own interest.

should be accompanied by minimal and distributional considerations of objective measures of well-being such as real opportunities or capabilities.

With respect to the second, the case is that the freedom or the opportunities one has may be valued *for themselves*. For instance, one may prefer a leader to be selected through an election rather than through appointment, irrespective of the identity of the leader that is finally chosen. Freedom in this sense is an end in itself. Neither first- nor second-order preferences over alternatives take this intrinsic value of opportunities into account. A potential escape to this limitation would be to follow Gravel (1994, 1998), who defines first-order preferences on the set of alternatives and on the opportunity sets that contain them, such that preferences are defined over some *extended* set of ordered pairs like (a, A) or $(a, \{a\})$. This allows to express statements such as "I prefer choosing a from set A to choosing the same a from set $\{a\}$" (Gravel 1994, 455). Second-order preference would be then defined over a pair of preferences, indicating if an individual identifies or not with her preferences over alternatives *and* her preferences over opportunity sets. An open question is if using second-order preferences would be useful to overcome some of the difficulties identified by Gravel (1994, 1998) with finding an ordering of the extended set of ordered pairs that respects individual preferences.

8.5 Hierarchical Models and the Inner Rational Agent

A question that may emerge from the reading of this Chapter is if adopting hierarchical models is not pushing the inner rational agent model one level up. In fact, when meta-preferences are discussed in economics they are often (implicitly or explicitly) assumed to be a stable and context-independent single ranking of multiple preferences. This corresponds, in some sense, to an inner rational agent model at the second-order level. The same is true if one assumes the existence of consistent *latent* second-order preferences which can be reconstructed by eliminating objective mistakes.

I would like to argue that the view exposed in this Chapter, in particular through the two hierarchical preferences models, is not necessarily in line with the inner rational view of agency. In the retrospective model, second-order preferences (both identification and valuation) are not assumed to be neither transitive nor complete (though they are assumed to be acyclic), and most importantly, they are assumed to be the result of a reflexive activity in *one period*. This means that the model does not impose stability of second-order preferences. I have neither assumed that second-order preferences are context-independent. In the evolving model, the agent is assumed to have the reflexive ability to evaluate his preferences/choices in several instances, but second-order preferences are again not assumed to be rational in terms of transitivity, completeness, and most importantly, stability or context-independence. Finally, all preferences (choices) are self-authenticated and second-order preferences cannot be judged mistaken from the *outside* within these models. This seems to contrast with most versions of the inner (and outer) rational agents found in the literature.

This of course is at some cost. The limits in terms of behavioral implications and welfare analysis of a model in which preferences and higher-order attitudes may be context-dependent and/or evolve over time is an open (and in many ways empirical) question. But one would certainly lose some parsimony and power in terms of prediction of behavior. In terms of welfare analysis, we have seen that more structure is needed in order to *guarantee* an acyclic welfare ranking that takes choices and reflexive preferences into account. Similarly, context-dependence upon the order at which stated identification preferences are elicited, could pose a problem for the use of such information in welfare analysis. In sum, I believe that many challenges face someone who wishes to take the time and context-dependency of preferences and higher-order attitudes seriously into consideration. But some strategies may be available, such as finding "adaptation and information proof" contexts to elicit first- and second-order preferences. In such settings, one could be more confident on fully respecting the preferences and judgments of individuals and on finding coherent and/or acyclic welfare rankings even though acyclicity would not be theoretically guaranteed.

9 Concluding Remarks

From this analysis, it seems that economics could gain from adopting a richer conception of the economic agent that accounts for some of the essential capacities that define a *person*, such as the ability to evaluate and change one's preferences, one's personal identity, or *who* one is. A person, according to the view exposed here, can either identify or not with her preferences and her preference change. Among other advantages, considering a *person instead of an agent* could get economics closer to what people care about, their goals, who they want to be or become, and what is important to them.

One can contrast this view with Sugden's (2004, 2007) proposal to model the economic agent as a "continuing agent", that can be seen "as the *composition* of the series of *time-slice agents*" (i.e., the composition of the agent in period 1, the agent in period 2, and so on) [see 2007, 671]. According to this view, an agent's choice in period t is not only interpreted as the deliberate choice of the "agent in period t" but also as the deliberate choice of the "continuing agent". Sugden (2004, 2007) sees this continuing agent as a continuing *locus of responsibility* and argues that this agent "*identifies* with each of his time-slices" (see 2007, footnote 5). What my arguments suggest is that there are many instances when people do not identify with what they have done or are about to do, and that this provides meaningful information to understand the nature of individuals' decisions.

Several questions are left open regarding how these notions relate with well-being, justice, and moral responsibility. Take the example of Jack, a husband that has betrayed his wife although he would have liked to remain faithful. According to his second-order preferences, Jack seems not to identify with his time-slice that betrayed his wife. Is this an indication that he regrets his (free) choice, and that this action has decreased his hedonic well-being? Is it fair, say for questions of "punishments and/or rewards", to differentiate a case like Jack's from that of a husband that identifies with his betrayal? Should the fact that Jack does not identify with his action excuse him of any or some moral responsibility? These questions illustrate the fascinating topics that are left to explore.

Finally, I have taken a deterministic view of agency but non-deterministic views of agency such as *random preferences models*, according to which agents' preferences change stochastically (e.g. Becker et al. 1963; Barberà and Pattanaik 1986; McFadden and Richter 1990; Loomes and Sugden 1995; Gul and Pesendorfer 2006; Apesteguia et al. 2017), or *deliberate randomization models*, according to which agents deliberately choose stochastically following a preference to reduce regret, incomplete preferences, difficulty to judge one's true risk aversion, and the like (e.g. Machina 1985; Marley 1997; Fudenberg and Strzalecki 2014), also rationalize (and predict) preference change.[47] These models could explain and rationalize, for example, that *ceteris paribus* an individual is more likely to change behavior when her first and second-order preferences conflict than when they don't. They are also alternative deliberate rationalizations of choices that are often judged in economics as mistakes. In fact, in a recent experiment with repeated choices on similar lotteries, Agranov and Ortoleva (2017) disentangle if stochastic choice (i.e., different choices when choosing from the same set of alternatives many times) can be rationalized by these models or by subjects' mistakes. Their main finding is that the majority of subjects choose stochastically on purpose rather than commit mistakes. This provides yet another argument in favor of not interpreting, *a priori*, any deviation of "rational" behavior as a mistake.

Acknowledgements I wish to thank Nicolas Gravel for his extensive and valuable comments.

References

Afriat, S. (1967) The Construction of Utility Functions from Expenditure Data. International Economic Review 8(1): 67–77.

Agranov, Marina and Pietro Ortoleva (2017) Stochastic Choice and Preferences for Randomization. Journal of Political Economy 125(1): 40–68.

Aizerman, M. A. and A. V. Malishevski (1981) General Theory of Best Variants Choice: Some Aspects. IEEE Transactions on Automatic Control 26(5): 1030–41.

Akerlof, G. A. (1991) Procrastination and Obedience. The American Economic Review 81(2): 1–19.

Apesteguia, J. and M. A. Ballester (2010) The Computational Complexity of Rationalizing Behavior. Journal of Mathematical Economics 46: 356–63.

——— (2015) A Measure of Rationality and Welfare. Journal of Political Economy 123(6): 1278–1310.

Apesteguia, J., M. A. Ballester, and J. Lu (2017) Single Crossing Random Utility Models. Econometrica 85(2): 661–74.

Arlegi, R. and M. Teschl (2015) Conflicts in Decision Making. In: C. Binder, G. Codognato, M. Teschl, and Y. Xu (eds) Individual and Collective Choice and Social Welfare: Essays in Honor of Nick Baigent. Springer-Verlag Berlin Heidelberg, : 11–29.

[47] See Fishburn (1999) for an old but comprehensive review.

Arrow, K. J. (1951) Social Choice and Individual Values. Wiley, New York.

——— (1959) Rational Choice Functions and Orderings. Economica 26: 121–27.

Baigent, N. (1995) Behind the Veil of Preferences. Japanese Economic Review 46(1): 88–101.

Barberà, S. and P. K. Pattanaik (1986) Falmagne and the Rationalizability of Stochastic Choices in Terms of Random Orderings. Econometrica 54(3): 707–15.

Becker, G., M. DeGroot, and J. Marschak (1963) Stochastic Models of Choice Behavior. Behavioral Science 8: 41–55.

Benjamin, D. J., J. J. Choi, and A. J. Strickland (2010) Social Identity and Preferences. The American Economic Review 100(4): 1913–28.

Benjamin, D. J., O. Heffetz, M. S. Kimball, and A. Rees-Jones (2012) What Do You Think Would Make You Happier? What Do You Think You Would Choose?. The American Economic Review 102(5): 2083–110.

Benjamin, D. J., O. Heffetz, M. S. Kimball, and N. Szembrot (2014) Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference. The American Economic Review 104(9): 2698–735.

Berlin, I. (1969) Four Essays on Liberty. Oxford University Press., Oxford.

Bernheim, B. D. and A. Rangel (2004) Addiction and Cue-Triggered Decision Processes. The American Economic Review 94(5): 1558–90.

——— (2007) Toward Choice-Theoretic Foundations for Behavioral Welfare Economics. The American Economic Review: Papers and Proceedings 97(2): 464–70.

——— (2009) Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. The Quarterly Journal of Economics 124(1): 51–104.

Bossert, W., Y. Sprumont, and K. Suzumura (2005) Consistent Rationalizability. Economica 72: 185–200.

——— (2006) Rationalizability of Choice Functions on General Domains Without Full Transitivity. Social Choice and Welfare 27: 435–58.

Bowles, S. (1998) Endogenous Preferences: The Cultural Consequences of Markets and other Economic Institutions. Journal of Economic Literature 36: 75–111.

Bratman, M. E. (1996) Identification, Decision, and Treating as a Reason. Philosophical Topics 24(2): 1–18.

——— (2003) A Desire of One's Own. The Journal of Philosophy 100(5): 221–42.

Brennan, T. J. (1993) The Futility of Multiple Utility. Economics and Philosophy 9: 155–64.

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

Bykvist, K. (2003) The Moral Relevance of Past Preferences. In: H. Dyke (ed) Time and Ethics: Essays at the Intersection. Kluwer, Dordrecht, Holland: 115–36.

Carpenter, J. P. (2005) Endogenous Social Preferences. Review of Radical Political Economics 37(1): 63–84.

Cowen, T. (1993) The Scope and Limits of Preference Sovereignty. Economics and Philosophy 9: 253–69.

Davis, J. B. (2009) Identity and Individual Economic Agents: A Narrative Approach. Review of Social Economy 67(1): 71–94.

Decancq, K., M. Fleurbaey, and E. Schokkaert (2015) Happiness, Equivalent Incomes and Respect for Individual Preferences. Economica 82: 1082–106.

Dennett, D. (1991) Consciousness Explained. Little Brown & Co.

Dietrich, F. and C. List (2013) Where do Preferences Come From?. International Journal of Game Theory 42(3): 613–37.

——— (2016) Reason-based choice and context-dependence: An explanatory framework. Economics and Philosophy 32(2): 175–229.

Elster, J. (1982) Sour Grapes - Utilitarianism and the Genesis of Wants. In: A. K. Sen and B. Williams (eds) Utilitarianism and Beyond. Cambridge University Press, Cambridge: 219–38.

——— (1985) Introduction. In: J. Elster (ed) The Multiple Self. Cambridge University Press, Cambridge: 1–34.

Fehr, E. and K. Hoff (2011) Introduction: Tastes, Castes and Culture: The Influence of Society on Preferences. The Economic Journal 211: 396–412.

Fischer, J. M. (1999) Recent Work on Moral Responsibility. Ethics 110(1): 93–139.

Fishburn, P. (1999) Stochastic Utility. In: S. Barberà, P. Hammond, and C. Seidl (eds) Handbook of Utility Theory. Vol. 1 Principles Kluwer Academic Publishers, Dordrecht, Holland: 273–320.

Fleurbaey, M. and E. Schokkaert (2013) Behavioral Welfare Economics and Redistribution. American Economic Journal: Microeconomics 5(3): 180–205.

Foster, J. (1993) Notes on Effective Freedom. Mimeo, Vanderbilt University.

Frankfurt, H. G. (1971) Freedom of the Will and the Concept of a Person. The Journal of Philosophy 68(1): 5–20.

——— (1988) Identification and Wholeheartedness. In: The Importance of What We Care About: Philosophical Essays. Cambridge University Press.

——— (1992) The Faintest Passion. In: Proceedings and Addresses of the American Philosophical Association. 66 American Philosophical Association., Newark, Del.: 5–16.

——— (2009) The Reasons of Love. Princeton University Press, Princeton, N.J.

Fudenberg, D. and D. K. Levine (2006) A Dual Self Model of Impulse Control. The American Economic Review 96: 1449–76.

——— (2012) Timing and Self-Control. Econometrica 80(1): 1–42.

Fudenberg, D. and T. Strzalecki (2014) Recursive Stochastic Choice. , Mimeo, Harvard University.

Gallagher, S. (2000) Philosophical Conceptions of the Self: Implications for Cognitive Science. Trends in Cognitive Sciences 4(1): 14–21.

George, D. (1984) Meta-Preferences: Reconsidering Contemporary Notions of Free Choice. International Journal of Social Economics 11(3/4): 92–107.

Gravel, N. (1994) Can a Ranking of Opportunity Sets Attach an Intrinsic Importance to Freedom of Choice?. American Economic Review Papers and Proceedings 84: 454–58.

——— (1998) Ranking Opportunity Sets on the Basis of their Freedom of Choice and their Ability to Satisfy Preferences: A Difficulty. Social Choice and Welfare 15: 371–82.

——— (2008) What is Freedom?. In: Handbook of Economics and Ethics. Edward Edgar Publishing, London.

Gul, F. and W. Pesendorfer (2001) Temptation and Self-control. Econometrica 69(6): 1403–36.

——— (2004) Self-control and the Theory of Consumption. Econometrica 72(1): 119–58.

——— (2005) The Revealed Preference Theory of Changing Tastes. The Review of Economic Studies 72(2): 429–48.

——— (2006) Random Expected Utility. Econometrica 74(1): 121–46.

——— (2008) The Case for Mindless Economics. In: A. Caplin and A. Schotter (eds) The Foundations of Positive and Normative Economics. Oxford University Press, New York: 3–39.

Harsanyi, J. C. (1997) Utilities, Preferences, and Substantive Goods. Social Choice and Welfare 14: 129–45.

Hausman, D. M. (2012) Preference, Value, Choice, and Welfare. Cambridge University Press, New York.

——— (2013) A reply to Lehtinen, Teschl and Pattanaik. Journal of Economic Methodology 20(2): 219–23.

Hirschman, A. O. (1984) Against Parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse. The American Economic Review: Papers and Proceedings 74(2): 89–96.

Hoff, K. and J. E. Stiglitz (2016) Striving for Balance in Economics: Towards a Theory of the Social Determination of Behavior. Journal of Economic Behavior and Organization 126: 25–57.

Horst, U., A. Kirman, and M. Teschl (2006) Changing Identity: The Emergence of Social Groups. GREQAM Working Paper.

Infante, G., G. Lecouteux, and R. Sugden (2016) Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics. Journal of Economic Methodology 23(1): 1–25.

Jamison, J. and J. Wegener (2010) Multiple Selves in Intertemporal Choice. Journal of Economic Psychology 31: 832–39.

Jeffrey, R. C. (1974) Preferences Among Preferences. The Journal of Philosophy 71(13): 377–91.

Jones, P. and R. Sugden (1982) Evaluating Choice. International Review of Law and Economics 2: 47–65.

Kahneman, D. (2003) Maps of Bounded Rationality: Psychology for Behavioral Economics. The American Economic Review 93(5): 1149–75.

——— (2011) Thinking, Fast and Slow. Farrar, Straus & Giroux, New York, NY.

Kahneman, D. and R. H. Thaler (2006) Anomalies: Utility Maximization and Experienced Utility. The Journal of Economic Perspectives 20(1): 221–34.

Kalai, G., A. Rubinstein, and R. Spiegler (2002) Rationalizing Choice Function by Multiple Rationales. Econometrica 70(6): 2481–88.

Kirman, A. and M. Teschl (2006) Searching for Identity in the Capability Space. Journal of Economic Methodology 13(3): 299–325.

Koszegi, B. and M. Rabin (2007) Mistakes in Choice-based Welfare Analysis. American Economic Review Papers and Proceedings 97(2): 477–81.

Laibson, D. (1994) Essays in Hyperbolic Discounting. Ph.D. dissertation, MIT.

——— (1997) Golden Eggs and Hyperbolic Discounting. The Quarterly Journal of Economics 112(2): 443–77.

LeBoeuf, R. A., E. Shafir, and J. B. Bayuk (2010) The Conflicting Choices of Alternating Selves. Organizational Behavior and Human Decision Processes 111(1): 48–61.

Lehtinen, A. (2012) A Review on Daniel Hausman (2012): Preference, Value, Choice, and Welfare.

Lewis, D. (1989) Dispositional Theories of Value. Proceedings of the Aristotelian Society, Supplementary Volumes 63: 113–137.

Lippert-Rasmussen, K. (2003) Identification and Responsability. Ethical Theory and Moral Practice 6: 349–76.

Livet, P. (2006) Identities, Capabilities and Revisions. Journal of Economic Methodology 13(3): 327–48.

Loomes, G. and R. Sugden (1995) Incorporating a Stochastic Element Into Decision Theories. European Economic Review 39: 641–48.

Machina, M. J. (1985) Stochastic Choice Functions Generated from Deterministic Preferences over Lotteries. The Economic Journal 95: 575–94.

Manzini, P. and M. Mariotti (2007) Sequentially Rationalizable Choice. The American Economic Review 97(5): 1824–39.

Marley, A. (1997) Probabilistic Choice as a Consequence of Nonlinear (sub) Optimization. Journal of Mathematical Psychology 41: 382–91.

McFadden, D. and M. K. Richter (1990) Stochastic Rationality and Revealed Stochastic Preference. In: J. S. Chipman, D. McFadden, and M. K. Richter (eds) Preferences, Uncertainty, and Optimality: Essays in Honor of Leo Hurwicz. Westview Press: Boulder, CO, 161-186., Boulder, Colorado: 163–186.

Nehring, K. and C. Puppe (1999) On the Multi-preference Approach to Evaluating Opportunities. Social Choice and Welfare 16: 41–63.

O'Donoghue, T. and M. Rabin (1999) Doing It Now or Later. The American Economic Review 89(1): 103–24.

——— (2001) Choice and Procrastination. The Quarterly Journal of Economics 116(1): 121–60.

——— (2003) Studying Optimal Paternalism, Illustrated with a Model of Sin Taxes. The American Economic Review: Papers and Proceedings 93(2): 186–91.

Parfit, D. (1984) Reasons and Persons. Oxford University Press, Oxford.

Pattanaik, P. K. and Y. Xu (1990) On Ranking Opportunity Sets in Terms of Freedom of Choice. Recherches Economiques de Louvain 56: 383–90.

Rabin, M. (2013) Incorporating Limited Rationality into Economics. Journal of Economic Literature 51(2): 528–43.

Ricoeur, P. (1984) Time and Narrative (3 Vols). University of Chicago Press.

——— (2002) Narrative Identity. In: D. Wood (ed) On Paul Ricoeur: Narrative and Interpretation. Routledge, London.

Rubinstein, A. and Y. Salant (2012) Eliciting Welfare Preferences from Behavioural Data Sets. The Review of Economic Studies 79: 375–87.

Schelling, T. (1984) Choice and Consequence: Perspectives of an Errant Economist. Harvard University Press, Cambridge, MA.

Schkade, D. A. and D. Kahneman (1998) Does Living in California Make People Happy? A Focusing Illusion in Judgments of Life Satisfaction. Psychological Science 9(5): 340–46.

Schotter, A. (2008) What's so Informative About Choice?. In: A. Caplin and A. Schotter (eds) The Foundations of Positive and Normative Economics. Oxford University Press, New York: 70–94.

Sen, A. K. (1971) Choice Functions and Revealed Preferences. Review of Economic Studies 38: 307–17.

——— (1974) Choice, Orderings, and Morality. In: S. Koerner (ed) Practical Reason, Oxford: Basil Blackwell.. Basil Blackwell, Oxford.

——— (1977) Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. Philosophy and Public Affairs 6(4): 317–44.

——— (1980) Plural Utility. In: Proceedings of the Aristotelian Society, New Series. 81 Wiley on behalf of The Aristotelian Society, : 193–215.

——— (1987) Commodities and Capabilities. Oxford University Press, New Delhi, Indiaoxford india paperbacks 1999 edition.

——— (1988) Freedom of Choice: Concept and Content. European Economic Review 32: 269–94.

——— (1991) Welfare, Preference and Freedom. Journal of Econometrics 50: 15–29.

Stigler, G. J. and G. S. Becker (1977) De Gustibus Non Est Disputandum. The American Economic Review 67(2): 76–90.

Stone, A. A., J. E. Broderick, L. S. Porter, and A. T. Kaell (1997) The Experience of Rheumatoid Arthritis Pain and Fatigue: Examining Momentary Reports and Correlates over one Week. Arthrities & Rheumatology 10(3): 185–93.

Stone, A. A. and S. Shiffman (1994) Ecological Momentary Assessment (EMA) in Behavioral Medicine. Annals of Behavioral Medicine 16: 199–202.

Strotz, R. H. (1955) Myopia and Inconsistency in Dynamic Utility Maximization. Review of Economic Studies 23(3): 165–80.

Sugden, R. (2004) The Opportunity Criterion: Consumer Sovereignty Without the Assumption of Coherent Preferences. The American Economic Review 94(4): 1014–33.

——— (2007) The Value of Opportunities over Time when Preferences are Unstable. Social Choice and Welfare 29: 665–82.

Thaler, R. H. and H. M Shefrin (1981) An Economic Theory of Self Control. Journal of Political Economy 89(2): 392–406.

Varian, H. (2006) Revealed Preferences. In: M. Szenberg L. Ramrattan and A. Gottesman (eds) Samuelsonian Economics and the Twenty First Century. Oxford University Press, Oxford: 99–115.

Watson, G. (1975) Free Agency. The Journal of Philosophy 72: 205–20.