Normative Economics and Paternalism: The Problem with the Preference-Satisfaction Account of Welfare

Cyril Hédoin[*]

Economics and management research center REGARDS – University of Reims Champagne-Ardenne, France

This version: 03/03/2015 – Do not quote without permission.

Abstract: The normative turn of behavioral economics has lead to a reconsideration of paternalism in normative economics. This article argues however that the preference-satisfaction account of welfare that still dominates welfare economics makes impossible to account for all the dimensions of the debate over paternalism. The laundered preferences approach and the alternative selves approach are two available frameworks to reconcile the consumer sovereignty principle that underlies the preference-satisfaction account with the fact that preferences are endogenous and context-dependent. I show however that neither of them is able to account for autonomy-related issues which are central in current debates over "soft" or "libertarian" paternalism. I suggest that a justification of paternalism compatible with liberal principles depends on the ability for reasonable persons to voluntarily consent to a collective choice rule with paternalistic tendencies. This argument relies on a distinction between preferences (which can be attached to other entities than persons) and values which is unknown to welfare economics.

Keywords: Normative economics – Paternalism – Preference-Satisfaction Account – Autonomy – Values – Behavioral Economics

0. Introduction

Welfare economics has been historically built on a more or less loosely defined principle of consumer sovereignty. This principle is closely articulated to the definition of welfare as preference-satisfaction that has dominated the "new" welfare economics since the pioneering work of Vilfredo Pareto, John Hicks and Nicholas Kaldor. Relying on several experimental results coming from behavioral economics, an increasing number of economists are however subscribing to paternalistic social evaluations and policies. Most of them are endorsing a "light" (Loewenstein and Haisley 2008) or "libertarian" (Sunstein and Thaler 2003) form of

[*] Full professor of economics. Contact: cyril.hedoin@free.fr
Paper prepared for the "Paternalism redeemed: old ideas, new realities" workshop, 30-31/03/2015, Lyon, France.

paternalism in an attempt to reconcile the libertarian foundations of normative economics with the empirical evidence that people often do not behave in their own interest.

An interesting feature of these normative extensions of behavioral economics is that most of them have not given up the welfarist criterion of preference-satisfaction. This is explicit in Douglas Bernheim and Antonio Rangel's significant attempt to build a "behavioral welfare economics" [(Bernheim and Rangel 2007); (Bernheim and Rangel 2009)]. This is also the case – though more implicitly – in the oft-cited writings of Thaler and Sunstein on libertarian paternalism. This article precisely deals with the philosophical and theoretical implications of this rather unnatural association between the welfarist criterion of preference-satisfaction and the rising paternalistic stance of a part of normative economics. I shall argue that the perspectives of this encounter between welfare economics and paternalism are not promising. The main reason lies in the fact that all the plausible conceptual and theoretical strategies to deal with the fact that preference are endogenous and more particularly context-dependent are ill-qualified to account for the issue of autonomy that is at the core of all discussions about paternalism – which is even more true when one is endorsing a "libertarian" or "soft" version of paternalism.

The fact that welfarism is unable to deal adequately with autonomy-related issues should not be a surprise. Amartya Sen's famous Paretian-liberal paradox (Sen 1970a) is an early result stating that a weak form of welfarism, Paretianism, is inconsistent with the satisfaction of a condition of minimal liberty. However, the normative turn of behavioral economics threatens the foundations of welfare economics even more deeply because it casts doubts on the very definition of welfare and on how we should measure it. The main point I shall make is that rescuing the preference-satisfaction welfare criterion against behavioral anomalies almost inevitably implies to give up the notion that a person is an economic agent and the main locus for welfare considerations. Since autonomy and freedom are values that necessarily pertain to persons, behaviorally informed welfare economics totally disconnects welfare and autonomy. As a result, welfare economics cannot be of any help to discuss paternalistic social evaluations and policies. Instead, following Gerald Dworkin [(1972); (2010)] and John Rawls (1971), I suggest that a justification of paternalism compatible with liberal principles depends on the ability for reasonable persons to voluntarily consent to a collective choice rule with paternalistic tendencies. This argument relies on a distinction between *preferences* (which can be attached to other entities than persons) and *values* unknown to welfare economics.

The article is divided into six sections. The first section discusses the general problem of endogenous preferences in normative economics and articulates it with experimental findings in behavioral economics. Section two considers the so-called "laundered preferences" or "informed-desires" approach to deal with endogenous preferences. As I note, normative economists have endorsed this approach even before the rise of behavioral economics. I underline however several defects that make it unable to deal with issues related to the autonomy of the person. Section three considers what I take to be a more promising theoretical strategy based on the notion of "multiple selves". I argue that Bernheim and Rangel's behavioral welfare economics in essence implements this strategy. I propose a formulation in terms of social welfare functional that states that paternalism can be

understood in terms of weighing across selves. The fourth section points to an obvious defects of this approach, namely that the person is no longer an economic agent and that therefore considerations of autonomy are meaningless. The fifth section argues that while preferences may be attributed to selves (or groups), *values* are only owned by persons. A plausible version (if any) of paternalism compatible with liberal principles depends on the ability of reasonable persons to recognize their behavioral failures and to agree over a collective decision rule that may not respect their preferences. The sixth section concludes by discussing some implications for the "nudge" approach defended by Thaler and Sunstein.

1. Normative Economics and Endogenous Preferences

Several theories of welfare (or well-being) have been competing against each other in philosophy and social sciences for several decades (and arguably, centuries). One may distinguish three generic accounts of welfare (Angner 2015): mental-state accounts, objective-list accounts and preference-satisfaction accounts. The first two correspond to what can be called "substantive" theories of welfare according to which "things are intrinsically good for people" (Hausman and McPherson 2006, 119). The latter rather correspond to "formal" theories which "specify how one finds out what things are intrinsically good for people, but they do not say what those things are" (Hausman and McPherson 2006, 119). Standard welfare economics and more generally normative economics as a whole are deeply committed to this latter view. Virtually all welfare economists take for granted that welfare can be measured as the degree of satisfaction of the agents' preferences – at least at the theoretical level. Combined with a welfare criterion (*e.g.* the Pareto criterion or the Kaldor-Hicks compensation criterion), this account makes possible to rank states of the world or social alternatives according to the amount (and possibly the distribution) of welfare in the population.

The preference-satisfaction account of welfare has been justified in economics on the basis of several arguments. A first one, which can be called the "epistemic argument", is that the individual knows better than everyone else what is good for her. Since any person has a privileged knowledge of what is intrinsically valuable to her, her preferences are assumed to reflect her welfare. The "pragmatic argument" is similar. It states that in practice, satisfying one's preferences will most of the time be in one's interest and thus enhance her welfare. The reason may be that the person knows better than everyone else what is good for her (in which case the epistemic and the pragmatic arguments overlap); alternatively, the explanation may be that empirically there is a strong correlation between what a person wants and what is good for her. A third argument is ethical and relies on the so-called principle of consumer sovereignty. According to it, it is morally best to make each person free to decide as she wants according to what she conceives to be best for her. These three arguments have not always been clearly distinguished and it is not easy to know which one of them really underlies the theorems of welfare economics.[1] Nevertheless, the consumer sovereignty principle is not

[1] Actually, there is a fourth possibility which I have not mentioned since it has little analytical interest: one can argue that welfare is nothing but preference satisfaction *by definition*. This seems to be Gul and Pesendorfer's

affected by the experimental results of behavioral economics – contrary to the epistemic and the pragmatic arguments – and the economist's willingness (including some behavioral economists) to continue to rely on the preference-satisfaction account indicates its centrality in welfare economics.

Standard welfare economics mostly relies on a revealed-preference framework. Combined with the preference-satisfaction account, this leads to the postulate that welfare is reflected in the choice each person makes or would make in a given situation. It follows that the ability to measure welfare, or at least to compare two social alternatives on a welfare basis, depends on the fact that people make *consistent* choices. Choice consistency can be defined on the basis of different axioms; a very weak one states that if one would choose alternative *x* when only alternatives *x* and *y* are available, then one should never choose *y* when both *x* and *y* are available along with other alternatives (Sugden 1985). This condition is weaker than all the usual consistency requirements used in consumer theory or social choice theory.[2] Under the assumption that choices reveal a binary relation *R* over any pair of alternatives, the stronger consistency requirements imply that the agents' preferences form a complete ordering; weaker requirements only guarantee preferences acyclicity.

Such consistency requirements are made for several reasons. From the point of view of positive economics, they make possible to infer from a finite set of observed choices unobserved choices over hypothetical decision problems. From the point of view of normative economics, the motivation is rather different. Given the preference-satisfaction account of welfare and the revealed-preference approach, choice consistency is required to make the welfare notion meaningful. Indeed, assume for instance that the minimal consistency requirement above is not satisfied and that some person *i* prefers *x* over *y* (*i.e.* chooses *x* rather than *y*) when only *x* and *y* are available, but chooses *y* when *x* and a third option *z* is also available. Then, it seems plainly impossible to state which of alternatives *x* or *y* is better in terms of *i*'s welfare. Since all welfare criteria of normative economics (in particular, the Pareto criterion) depend on our ability to make such evaluative statement, to give up choice consistency would also mean to give up of possibility to make welfare judgment.

The fact that the very existence of welfare economics depends on the consistency of choices and preferences is a reason why the preference-satisfaction account has been regarded with suspicion outside economics. In particular, the problem of *endogenous preferences* seems to threaten normative economics as a whole if the preference-satisfaction account is used to

---

(2008) view hiding behind their claim that welfare economics is actually about *positive* rather than normative economics. Then, welfare economics consists in a set of statements about the properties of some institutional setting that depend only on logical relations. This seems dubious because such a view depends on a strong dichotomy between (judgment of) facts and (judgments of) values but also because historically the theorems of welfare economics have been used as arguments for the superiority of market economies over other systems of resources allocation. Finally, let me note that if it was true, this view would completely disqualify cost-benefit analysis, since the latter makes normative judgments on the basis of people's willingness to pay, and thus indirectly on their preferences over alternatives.

[2] In consumer theory, choice consistency is most of the time defined on the basis of the "weak axiom of revealed preference" (Samuelson 1938) which states that if a bundle of goods *x* is chosen over a bundle *y* for some budget limit and price vector, then *y* will never be chosen over *x* when *x* is affordable. In social choice theory, the weaker "basic contraction axiom" is virtually always required. According to it, if *x* is chosen for a set *S* of available alternatives, then it must also be the case for any smaller set *S'* where *x* is available.

make welfare judgments. Endogenous preferences mean that preferences – and the choices they are revealed from – change as a function of some variables not directly under the control of the decision maker. Variables responsible for such a change can be the time at which a choice is made, the context in which the decision maker is embedded or choices made in the past.[3] Jon Elster (1985) and Amartya Sen (1991) have been early critics of the preference-satisfaction account on the basis of the fact that preferences are "adaptive", *i.e.* that individuals tend to adapt their preferences to the circumstances in which they are used to live. Similarly, Tyler Cowen (1993) convincingly argues that endogenous preferences impose a serious limitation upon the principle of consumer sovereignty and thus make the preference-satisfaction account of welfare far less attractive. As Cowen notes, endogenous preferences are particularly problematic for welfare economics "when preferences (or metapreferences) are determined by the policies being chosen or evaluated" (p. 254). Indeed, in this case, policy evaluations will not necessarily lead to the same result whether we measure welfare on the basis of *ex ante* or *ex post* preferences. This problem is formally identical to the so-called "Scitovsky double-switching problem" in cost-benefit analysis. According to the latter, in the case we use *ex post* preferences a social alternative A could be welfare-superior to an alternative B if the status quo is B, and B welfare-superior to A if the status quo is A, because of the wealth effects generated by the policy chosen. However, the problem posed by endogenous preferences is even more general: not only they can lead to the "sour grapes" phenomenon described by Elster (1985) where to institute the social alternative A creates an environment that leads individuals to prefer social alternative B (and vice-versa), they can also make welfare analysis meaningless even when only one individual is concerned (Cowen 1993, 257).

Clearly, the problem of endogenous preferences has been recognized well before the emergence of behavioral economics. It also can plausibly lead to the recommendation of some form of paternalism (Qizilbash 2009), independently of the experimental evidence provided by behavioral economists about people's rationality. Still, behavioral economics has made a huge contribution from this perspective because it shows the extent to which preferences are *context-dependent*. Consider for instance the following three well-established behavioral "anomalies" largely documented in the literature: hyperbolic discounting and preference reversal, endowment effect, framing effect. In the case of hyperbolic discounting (Strotz 1955), people value future gains differently as a function of the point in time they make their valuation. More exactly, the relative valuation of the sooner rewards compared with the later rewards increases the shorter is the delay for receiving the former. Hyperbolic discounting then leads to intertemporal inconsistencies because preferences over two kinds of rewards

[3] In the latter case, changing preferences can be accommodated by postulating that preferences change on the basis of past choices according to some "metapreferences". This is the approach famously endorsed by Stigler and Becker (1977). However, this approach solves the problem of endogenous preferences in normative economics only provided that agents are able to foreseen the consequences of their choices on the content of their future preferences. This a strong assumption to say the least. Clearly, behavioral economics makes it even less plausible.

(sooner and smaller ones versus later and larger ones) change as time is passing by. Here, time is the contextual element that makes preferences endogenous.[4]

The endowment effect is particularly relevant in the case of cost-benefit analysis because it indicates that an individual's willingness to pay and willingness to accept some good or policy may differ. The point is that one's valuation of some good or some state of affair differs depending on whether or not she has the "rights" over this good or state. In this case, an agent's preferences depend on the initial allocation of rights. Finally, the framing effect can be seen as a generalized statement for the fact that preferences are context-dependent. It states that the way a decision problem is framed to the decision maker may change the preferences the latter reveals through her choice. Crucially, the framing of the decision problem is irrelevant because it does not alter the problem's nature and content but nevertheless is sufficient the change one's behavior. Here, preferences simply change with the framing used. Hyperbolic discounting, the endowment effect and the framing effect thus all establish that preferences are highly sensitive to several "non essential" variables. This can be considered to be the key insight of behavioral economics to the problem of endogenous preferences : not only are preferences changing as a function of a set of variables not under the control of the decision maker, these variables (or at least some of them) are ostensibly *welfare-irrelevant*. Moreover, they may lead the individual to reveal inconsistent preferences. Not only this threatens the very possibility to identify welfare to preference-satisfaction (see above); this also indicates that at least in some cases individuals may be induced to make choices that contradict previously revealed preferences which, we assume, are a basis to make welfare judgment. Now, if one insists on keeping with the preference-satisfaction account of welfare on the ground of the consumer sovereignty principle, some adjustments to account for endogenous preferences are clearly required. These adjustments however inevitably have paternalistic tendencies which introduce into welfare economics a new kind of considerations related to autonomy and freedom.

2. Paternalism and the "Laundered Preferences" View of Welfare

A first way to deal with endogenous preferences in the preference-satisfaction account consists in equating welfare with the satisfaction not of *actual* preferences but rather of *informed* or *laundered* preferences. The point is that while actual preferences may endogenously change as a function of welfare-irrelevant variables, informed preferences are presumed to be stable. The claim that welfare must be defined as the satisfaction of informed preferences actually largely predates the experimental results of behavioral economics. It is

[4] We should distinguish the case of preference reversal due to hyperbolic discounting from behavioral inconsistencies due to strategic inconsistency and learning effects with imperfect information, even though they may take an intertemporal form. The latter two do not involve an authentic preference reversal since inconsistent behavior is generated by a strategic or an informational element that was not foreseen by the agent at the moment she makes her first decision. No preference change is involved in this case, unless we stick to a purely revealed-preference framework where preferences are defined as choices. While strategic inconsistency may be qualified as "irrational", it is not so clear in the learning case. However, we may speak of irrationality if the individual fails to make proper use of a new information, thus leading to "imperfectly informed preferences" (Cowen 1993). But such imperfectly informed preferences are not due to hyperbolic discounting.

not necessarily related to the problem of endogenous preferences either. For instance, in his writings defending a preference-based version of utilitarianism, the economist John Harsanyi claimed that not all preferences should be regarded as welfare-relevant:

> "It is, of course, well known that a person's preferences may be distorted by factual errors, ignorance, careless thinking, rash judgments, or strong emotions hindering rational choice, etc. Therefore, we may distinguish between a person's *explicit* preferences, i.e., his preferences as they actually *are*, possibly distorted by factual and logical errors – and his 'true' preferences, i.e., his preferences as they *would* be under 'ideal conditions' and, in particular, after careful reflection and in possession of all relevant information. In order to exclude the influence of irrational preferences, all we have to do is to define social utility in terms of the various individuals' 'true' preferences, rather than in terms of their explicit preferences." (Harsanyi 1977, 29-30).

Harsanyi went even further, suggesting that "[w]e have to disregard, not only preferences distorted by factual or logical errors, but also preferences based on clearly antisocial attitudes, such as sadism, resentment, or malice" (Harsanyi 1977, 30).[5] The idea that only informed or rational preferences should be taken into account when making welfare judgments is thus hardly new in normative economics. However, the experimental results of behavioral economics seem to make this approach even more appealing. Consider for instance the framing effect. Experimental findings tend to establish that people's choices are inconsistent across frames despite the fact that the decision problem remains the same. Not only this indicates that people's preferences are influenced by welfare-irrelevant factors; this also points out that at least some of these choices fail to reveal preferences that can be meaningfully related to the persons' welfare. This is a specific manifestation of the more general problem of endogenous preferences: unless one is ready to concede that welfare is itself an inconsistent and meaningless notion, preferences changes across different context imply that the satisfaction of some preferences cannot enhance one's welfare. Once again, to compare two or several alternatives on the basis of the degree of satisfaction of the persons' preferences, welfare judgments need to be based on preferences that are invariant across these alternatives. Otherwise, it is impossible (or meaningless) to say that some alternative is better than another one in terms of welfare. The framing effect only strengthens this problem by showing that preferences may change even between two situations that are identical except for some "ornamental" features.

Therefore, it is not really surprising that several behavioral economists have recently endorsed the laundered preference approach of welfare when discussing the normative implications of their experimental findings. The most significant example is provided by Thaler and Sunstein's discussion of nudges (Thaler and Sunstein 2009) and their defense of "libertarian paternalism" (Sunstein and Thaler 2003). Starting from the fact that "in many domains, people lack clear, stable, or well-ordered preferences", these authors argue for "the possibility that in some cases individuals make inferior decisions in terms of their own welfare –

[5] In a later writing, Harsanyi (1996) continued to argue that welfare should be defined as the satisfaction of informed preferences. However, he was also claiming that humans' basic desires are almost all the same, which makes his view very similar to an objective-list account. Indeed, he dared to propose such a list (Harsanyi 1996, 139).

decisions they would change if they had complete information, unlimited cognitive abilities, and no lack of self-control" (Sunstein and Thaler 2003, 1161-2, my emphasis). This suggests that Sunstein and Thaler are implicitly endorsing a version of the laundered preference approach of welfare.[6] Clearly, this approach involves a departure from the traditional revealed-preference framework that is still dominant in welfare economics. Welfare is identified to preferences that *would be* revealed in very specific and ideal circumstances. This creates a tension with the consumer sovereignty principles that underlies the preference-satisfaction account because the possibility that individuals do not choose in their own interest is recognized. Hence the paternalistic stance of Sunstein and Thaler's normative views.

At the same time, Sunstein and Thaler's endorsement of the laundered preference approach indicates that they still adhere to the preference-satisfaction account of welfare. Moreover, they do not completely reject the consumer sovereignty principle. Firstly, they commit themselves to a "libertarian" form of paternalism. Secondly, as suggested by Qizilbash (2012), it is possible to relate Thaler and Sunstein's laundered preference approach to Peter Railton's ideal adviser account of well-being (Railton 1986). On Railton's account, a person *A*'s rational preferences are the preferences that an ideal adviser *A+* endowed with "unqualified cognitive and imaginative powers, and full factual and nomological information about his physical and psychological constitution, capacities, circumstances, history" would like *A* to have (Railton 1986, 173). Crucially, *A+* needs not be someone else, only an idealized version of *A*. On this view, it could be argued that judging *A*'s welfare on the basis of *A+*'s views respects *A*'s sovereignty.

Nevertheless, the laundered preference approach has several defects. A first one is related to the content of rational preferences and their determination. A second one which is more relevant in the discussion about paternalism is about the view of the person that underlies this account. The difficulty to determine what are "rational" preferences is generally viewed as an important lacuna of the laundered preferences approach. In particular, even with Railton's ideal adviser account, it is not at all clear which theory of (epistemic and practical) rationality should be used to judge the rationality of preferences. In part, this is due to the fact that nobody is actually endowed with "unqualified cognitive and imaginative powers" and "full factual and nomological information about his physical and psychological constitution, capacities, circumstances, history". Unless one is ready to assume that the philosopher or the economist has superhuman cognitive abilities and uncommon factual and nomological knowledge, it is impossible to know for this epistemic reason what are the preferences *A+* would like *A* to have. Therefore, to ground paternalistic evaluations and policies on such kind of knowledge seems simply to be unrealistic or at least hardly attractive.

This critique applies to Sunstein and Thaler's defense of libertarian paternalism (Qizilbash 2012). It seems hard to know which decisions an individual with "complete information, unlimited cognitive abilities, and no lack of self-control" would make, particularly given the fact that behavioral economics has largely contributed to show that most if not all persons do not behave on the basis of these properties. However, Sunstein and Thaler's reliance on

[6] For similar statements, see Qizilbash (2012) and Sugden (2008).

behavioral economics suggests that the theory of practical and epistemic rationality they have in mind could be related to the axioms of expected utility theory that is at the core of modern economics since the 1950's. Indeed, most of the positive program of behavioral economics has consisted into evaluating the distance between the predictions of expected utility theory and the actual behavior of people in different settings. This is confirmed in a recent article by Sunstein (2012) who quotes Riccardo Rebonato's (2012) characterization of libertarian paternalism:

> "Libertarian paternalism is the set of interventions aimed at overcoming the unavoidable biases and decisional inadequacies of an individual by exploiting them in such a way as to influence her decisions (in an easily reversible manner) towards choices that she herself would make if she had at her disposal unlimited time and information, and the analytic abilities of a rational decision maker (more precisely, of Homo Economicus)" (Rebonato 2012, quoted in Sunstein 2012, 27).

Interestingly, while Sunstein claims that this definition is "imprecise" in "three respects", he does not reject the claim that libertarian paternalism assimilates the rational decision maker to the economist's *homo economicus* figure. The normative status of expected utility theory is explicitly assumed by Jose-Luis Pinto-Prades and Jose-Maria Abellan-Perpinan (2012) in a more applied perspective. Noting that "libertarian paternalists are not very specific about how to identify those situations where choices do not reveal true preferences" (p. 571), they propose a procedure to provide a quantitative estimate of the deviation of people's actual behavior from the rational one. It uses expected utility theory as "the right model for normative analysis" (p. 573) but assumes that people's behavior is best described by cumulative prospect theory. As the authors recognize, this procedure has a paternalistic stance: "The method we propose to estimate utilities is based on the assumption that there is a "true" utility that is not directly observed. What we observe is a biased estimate of this utility. It is biased because it is affected by probability distortion and loss aversion. We are then assuming that probability distortion and loss aversion are biases, that is, they are not normatively desirable" (p. 581).

The use of expected utility theory as a normative benchmark does not fully answer the first critique against the laundered preferences approach. An obvious difficulty is that the privileged normative status of expected utility theory is debatable. It is far from clear that the transitive axiom or the independence axiom should be components of the most relevant theory of practical rationality. Moreover, expected utility theory has nothing to say about the *content* of preferences; it only states that they must be consistent in some specific way. However, it seems that at least some of the nudges suggested by Sunstein and Thaler are at least as motivated by the fact that people pursue the "wrong" ends as by the fact that they are not able to reach their ends efficiently.[7] Even if we ignore these difficulties, a second critique remains valid: the laundered preference approach is probably unhelpful as a defense of paternalism

[7] See Sunstein's (2012) distinction between "means paternalists" and "ends paternalists". Sunstein claims that "[b]ehavioral economists generally focus on paternalism about means, not ends" (2012, 7). But he also recognizes that the "distinction between means and ends raises a number of difficult puzzles" (2012, 12). Actually, Sunstein (2012, 26) even reaches the conclusion that in some cases (*e.g.* problems of time-consistency) the distinction is not far from being meaningless.

because it cannot deal properly with the notion of autonomy. Indeed, most debates around paternalism involve a discussion of a principle formalized by John Stuart Mill in *On Liberty* (Mill 1859) and according to which one "cannot rightfully be compelled to do or forbear because it will be better for him to do so, because it will make him happier, because, in the opinion of others, to do so would be wise, or even right" (quoted in Dworkin 1972, 64). Mill's original argument against the paternalistic interferences with one's liberty is essentially consequentialist: "the strongest of all arguments against interference of the public with purely personal conduct is that when it does interference, the odds are that it interferes wrongly and in the wrong place" (Mill 1859, quoted in Dworkin 1972, 71). However, as Dworkin (1972) points out in his classical article on paternalism, there is also a non-consequentialist argument against paternalism that emerges from Mill's discussion: "It is because coercing a person for his own good denies this status as an independent entity that Mill objects to it so strongly and in such absolute terms" (Dworkin 1972, 74-5). In this perspective, paternalistic interferences are "justified only to preserve a wider range of freedom for the individual in question" (Dworkin 1972, 76).

As I argue above, given the experimental results of behavioral economics, it seems that the consumer sovereignty principle is the main reason why most economists still retain the preference-satisfaction account of welfare. The non-consequentialist argument against paternalism is also obviously taken seriously by Sunstein and Thaler, given their emphasis on the "libertarian" aspect of their proposal. Therefore, it seems hard if not impossible to justify paternalism on purely welfarist or consequentialist grounds; at least, the case for paternalism must recognize the status of the person as an "independent entity". The laundered preferences view of welfare seems however to deny this particular status to the person. The reason is suggested by Cowen (1993) in the specific case of intertemporal inconsistencies due to imperfectly informed preferences:

> "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).

Cowen's argument is not directed toward the issues of autonomy and is only indirectly related to the status of the person as an independent entity. Rather, it points out the fact that provided one's with better information may change one's preferences; however, Cowen argues, it is not clear why the welfare of the hypothetical better informed person should have special relevance. What matters is the welfare of the actual person because this is the only who exists. Though directed against the preference-satisfaction account, this point may be even stronger against mental-state accounts of welfare. But consider the following variant: by satisfying the preferences that an ideal adviser *A+* would like her actual counterpart *A* to have, we are not recognizing that *A* is an independent entity in her own rights. Unless *A* explicitly agrees to endorse the "rational" preferences as they are conceived by *A+*, by only taking into account of the rational preferences in the welfare analysis we are actually denying that *A* is the normative locus of the analysis. In other words, we are not really satisfying *A*'s preferences but rather

the preferences of someone else endowed with a different identity. This seems plainly incompatible with the consumer sovereignty principle.

3. Paternalism and Separability: The Alternative Selves Framework

In this section, I investigate an alternative way to reconcile the preference-satisfaction account of welfare with paternalistic tendencies in normative economics. It consists in separating the person into several *alternative selves*, each endowed with a consistent set of preferences and to weight differently these preferences into a social welfare function (or functional) on the basis of some normative criteria. Combined with some assumptions about the separability properties of the social welfare function, this permits to represent paternalistic evaluations in an additively separable form. Though this approach has not been explicitly formalized, I contend that it has strong affinities with Bernheim and Rangel's behavioral welfare economics. The main advantage of the alternative selves approach over the laundered preferences approach is that it only takes into account preferences of actual selves, not hypothetical ones.

So-called "multiple selves models" are increasingly popular in economics. Their development is tightly related to the theoretical and experimental insights provided by behavioral economics. These models are built to capture empirically significant behavioral "anomalies" or "failures", leading for instance to intertemporal inconsistency, by modeling the person as a community of agents with partially conflicting interests. Arguably, such a modeling device builds on a specific notion of economic agency where an agent is identified to a set of well-ordered preferences; this potentially has tremendous implications both for positive and normative economics [(Hédoin 2015); (Ross 2005)]. However, the notion of "alternative selves" was already used in normative economics before the emergence of behavioral economics. A significant example is provided by James Mirrlees' (1982) discussion of the status of utilitarianism in economics. Mirrlees notes that "[w]hat a person plans to do can be described as the totality of what he plans to do at particular times, and under particular circumstances" (Mirrlees 1982, 66). Mirrlees combines the preference-satisfaction account of welfare with the key notion of *separability*. He contemplates in particular the possibility to assign numerical utilities to a person's actions in each time-period and particular circumstance. This requires the person's preferences in each time-period and particular circumstance to be weakly separable, *i.e.* "it is necessary that his preferences regarding what he will be doing at one particular time in one particular set of circumstances be independent of what he may be planning for all other times and circumstances" (Mirrlees 1982, 66). Under the stronger condition of additive separability, where one's preferences in *any two states* (time-periods or sets of circumstances), taken together, are independent of his preferences in all other states, then person's preferences can be represented as the weighted sum of utilities over all the possible states.

Additive separability plays an essential role in the debates over utilitarianism in normative economics. For instance, John Harsanyi's "aggregation theorem" (Harsanyi 1955) provides an axiomatic defense of preference-based utilitarianism on the basis of separability conditions for

states of nature and persons.[8] Mirrlees' suggestion is to extend the use of separability conditions *within* the person, *i.e.* to capture the person's behavior as the result of independent selves each endowed with weakly or even additively separated and well-ordered preferences. From the normative point of view, it becomes now possible to define a person's welfare not as the satisfaction of all her preferences but rather as the satisfaction of *some* of them, *i.e.* to give different normative weights to the various alternative selves that constitute the person. Though they are using a very different, choice-based framework, I shall suggest that Bernheim and Rangel's behavioral welfare economics may support paternalistic evaluations and policies along similar lines.

In a series of articles, Bernheim and Rangel [(2007); (2008); (2009)] adapt the revealed-preference framework of standard welfare economics to account for the experimental finding that most of the time people fail to reveal consistent preferences. They explicitly note that their main task is to "review the standard perspective on individual welfare" (Bernheim and Rangel 2008, 157), *i.e.* to articulate the preference-satisfaction account with the possibility of inconsistent preferences. They retain most of the standard framework for describing choices and assessing individual welfare. Indeed, they develop a generalized version of the standard framework, where the latter is reduced to a special case. This is done by the introduction of the key notion of "generalized choice situation" (GCS) G which is defined as the extension of the standard notion of choice situation to the case where one's choice depends on features habitually considered as irrelevant. Denote $\boldsymbol{X}$ the set of objects or alternatives over which the individual as to choose and $X \subseteq \boldsymbol{X}$ some (sub)set of available alternatives. Finally, $\mathcal{X}$ is the set of all possible subsets X. In the standard framework, a choice situation is simply defined by X and $\mathcal{X}$ thus corresponds to the domain of choice situations that concerns welfare economics. On this basis, the choices an individual is making or would make are described by a choice function $C(.)$ defined as a functional relation C: $\mathcal{X} \rightarrow \boldsymbol{X}$, where we assume that $C(X) \subseteq X$ for all $X \subseteq \mathcal{X}$. $C(X)$ thus corresponds to the alternative(s) an individual would choose if only the objects in X were available.[9] Bernheim and Rangel's GCS is obtained by pairing a set of available alternatives X with an "ancillary condition" d. Thus, $G = (X, d)$ and we denote $\mathcal{G}$ the set of GCS which are considered to be relevant. I will assume here that $\mathcal{G}$ is defined as the Cartesian product of all the standard choice situations $\mathcal{X}$ and the set D of ancillary conditions, *i.e.* $\mathcal{G} = \mathcal{X} \times \mathrm{D}$.[10]

The notion of "ancillary conditions" is obviously essential here since it is responsible for the transformation of standard choice situations into generalized ones. An ancillary condition is "a feature of the choice environment that may affect behavior, but that is not taken to be a welfare-relevant characteristic of the chosen object" (Bernheim and Rangel 2008, 159). For example, the way an information or a decision problem is framed corresponds to an ancillary condition. Regarding intertemporal choice, the time at which the decision is made and more

[8] See Broome (1991) for a technical and philosophical discussion of Harsanyi's theorem which emphasizes the role of these separability conditions.

[9] It is also generally assumed that $C(X) \neq \varnothing$ for all $X \subseteq \mathcal{X}$.

[10] Note that Bernheim and Rangel do not make this assumption. This will be of importance below when I will compare Berhneim and Rangel's approach with an alternative selves framework.

generally the decision tree used are ancillary conditions. A key aspect of an ancillary condition is that they are welfare-irrelevant: the fact that choices are inconsistent across ancillary conditions cannot be meaningfully related to welfare considerations. Therefore, if choices reveal preferences, not all preferences are significant from the normative point of view. A straightforward illustration is the following: consider the following two sets of available alternatives $X = \{x, y\}$ and $X' = \{x, y, z\}$ and two possible ancillary conditions d' and d''. Choice inconsistency due to dependence over welfare-irrelevant variables results if we have $C(X, d') = x$ and $C(X', d'') = y$. Indeed, this is a violation of minimal choice consistency as defined above. More prosaically, it is also perfectly possible to have $C(X, d') = x$ and $C(X, d'') = y$.

Choice-inconsistency across ancillary conditions considerably restricts the scope of the welfare analysis. Depending on how we construct the binary relations representing weak preference, strong preference and indifference on the basis of the choice function $C(.)$, the preference ordering may fail to be complete or transitive.[11] This is also illustrated by the permissibility of Bernheim and Rangel's criterion of "individual welfare optima". Given my assumption that $\mathcal{G}$ is rectangular (*i.e.* $\mathcal{G} = \mathcal{X} \times D$), this criterion is equivalent to the more traditional "multi-self Pareto optima" criterion used in behavioral economics (Bernheim and Rangel 2008, 171). This leads to interpret each choice made under a particular ancillary condition d as the revealed preference R_d of some alternative self k. Consider the following definitions: an alternative x is a *strong* multi-self Pareto improvement over y if $x \in C(X, d)$ and $y \notin C(X, d)$ for any $X \in \mathcal{X}$ where $x \in X$ and for all $d \in D$. Then, x is a *weak* multi-self Pareto optimum if no strong multi-self Pareto improvement exists. The multi-self criterion allows identifying unambiguous welfare improvements for each person. However, it is clear that under choice-inconsistency across ancillary conditions, the set of weak multi-self Pareto optima may be fairly large and thus relatively unhelpful.[12]

Once we leave the individual level for the aggregate level, the welfare analysis is even more restricted. Consider two (social) alternatives x and y, where any social alternative is a complete description of all *welfare-relevant* features for the members of the population (income level, health state, …). In standard welfare economics, x and y will generally be ranked according to the (strong or weak) Pareto criterion or (particularly in cost-benefit analysis) to the Kaldor-Hicks compensation criterion. As this is well known, a major limitation of the Pareto criterion is that many social alternatives cannot be compared on this basis. However, once we allow for choice-inconsistency, massive incomparability will result: on the basis of a generalized version of the weak multi-self Pareto criterion, an alternative x is better than an alternative y if and only if x is a multi-self Pareto strong improvement over y for all persons. Clearly, the set of such "generalized" weak multi-self Pareto optima is at least as

[11] See Bernheim and Rangel (2008, p. 164-6). They show however that it is possible to construct an acyclic strict preference relation P^* and note that most of time acyclicity is a sufficient condition to conduct welfare analysis (since generally acyclicity guarantees the existence of maximal elements in the set X).

[12] The criterion of *strong* multi-self Pareto optimum is obtained in a similar fashion than the standard case. With choice-inconsistencies across ancillary conditions, no such optimum may exist.

large as the largest set of individual weak multi-self Pareto optima in the population and will be generally much larger.[13]

Bernheim and Rangel suggest a way to overcome this difficulty, by deleting irrelevant or suspect GCSs from the choice data set through the use of non-choice information and on the basis of some normative or evaluative criteria:

> "Thus, for example, if someone chooses x from X under condition d' where she is likely to be distracted, and chooses y from X under condition d'' where she is likely to be focused, we would delete the data associated with (X, d') before constructing the welfare relations. In effect, we take the position that (X, d'') is a better guide for the planner than (X, d'). Even with the deletion of choice data, these welfare relations may remain ambiguous in many cases due to other unresolved choice conflicts, but… the sets of weak individual welfare optima grow (weakly) smaller" (Bernheim and Rangel 2008, 186).

Interestingly, they suggest that "this refinement agenda entails only a mild modification of the core libertarian principles" since they "do not propose *substituting* nonchoice data, or any external judgment, for choice data", but rather "adhere to the principle that the social planner's objective should be to select an alternative that the individual would select for herself in *some* generalized choice situation" (Bernheim and Rangel 2008, 186-7, emphasis in original).

Before discussing further these points, let me take advantage of Bernheim and Ranger's suggestion to use non-choice information in the welfare analysis. Indeed, this makes possible to formalize the collective choice that results from paternalistic considerations in terms of a *social welfare functional*, *i.e.* a "mechanism that specifies one and only one social ordering given a set of individual welfare functions, one function for each individual" (Sen 1970b, 124). We will need to slightly change the notation. We now denote X as the set of social alternatives and we figure out which of these alternatives a benevolent planner would choose on the basis of some social welfare functional (SWFL), given the fact that people may exhibit inconsistent preferences. A SWFL will generate an ordering of the members of X on the basis of a vector of utility numbers (u_1, …, u_n), one per individual. This ordering may itself be represented by a utility function $U(.)$ unique up to any positive monotonic transformation. One obvious difficulty is that the utility numbers are meaningless unless each person's preferences satisfy minimal conditions of consistency such as transitivity; but the fact that preferences are context-dependent implies that generally this will not be the case. However, if each ancillary condition d activates a different alternative selves k for each person, then we can plausibly assume that each self k of any person i can be ascribed a utility function $u_{ik}(.)$. The SWFL is thus constructed on the basis of vector of utility numbers (u_{11}, u_{12}, …, u_{1m}, …, u_{nm}) for n persons and m alternative selves per person.

Extending Mirrless' (1982) suggestion to the interpersonal case, suppose that the ordering represented by $U(.)$ is strongly separable across selves, *i.e.* the preferences of any pair of

[13] Bernheim and Rangel (2008, 181-3) discusses this point using the Edgeworth box. When choices are inconsistent across ancillary conditions, the contract curve is thicker than in the standard case. The number of "generalized" weak Pareto optima is thus larger.

selves are independent from the preferences of all other selves. Then a theorem due to Gorman (1968) states that $U(.)$ is additively separable:

(1) $U(x) = \sum_i \sum_k \alpha_{ik} u_{ik}(x)$, for any $x \in X$.

Expression (1) makes sense only if the utility numbers are cardinally significant (at least on the basis of an interval scale) and comparable (at least in terms of utility differences). Then, (1) can be seen as an instantiation of a peculiar form of weighted utilitarianism.[14] The weighing coefficients $\{\alpha_{ik}\}$ correspond to the planner's normative or evaluative judgment over the welfare-relevance of the selves' preferences. Actually, to assume that the utility numbers are cardinally significant is plausible if we follow Bernheim and Rangel's suggestion to use non-choice information in the welfare analysis. This allows for the extension of the informational basis and thus to use more detailed utility information than in standard welfare economics. A straightforward implementation of Bernheim and Rangel's strategy of deletion of irrelevant GCSs simply consists into setting the coefficients $\alpha_{ik} = 0$ and to give a strictly positive value to other coefficients.[15]

We have thus arrived at an interesting formulation of paternalism in terms of SWFL. Formally, the SWFL is paternalistic if and only if 1) it is additively separable across alternative selves and 2) does not respect anonymity with respect to alternative selves. The resulting ordering then reflects a *paternalistic social evaluation*. Correspondingly, a *paternalistic social decision function* always selects the optimal available social alternative(s) such as identified by the paternalistic social evaluation. One virtue of this approach compared to laundered preference approach is that, as suggested by Bernheim and Rangel, the paternalistic planner is not trying to satisfy the hypothetical preferences of some idealized set of persons. The welfare analysis is still done on the basis of choice data and the selves that are given positive weight in the SWFL correspond to real choice patterns.[16] Thus, the critique of the previous section has no force here.

4. Reintroducing Autonomy: From Selves to Persons

An obvious objection to the approach of the preceding section is relative to the evaluative and normative criteria used for deleting GCS (or, equivalently, for setting the value of the weighing coefficients in the SWFL). Some behavioral and neuro- economists are probably counting on the advancements of neurosciences to help to deal with this issue in a scientific

[14] Expression (1) satisfies the two standard axioms of continuity and weak Pareto but it does not satisfy anonymity, which is an important axiom in the multi-profile approach in social choice theory (Blackorby, Bossert, and Donaldson 2005). This is due to the fact that the coefficients α_{ik} have not to be equal. Note moreover that (1) does not presuppose that each person i's utility function is additively separable, *i.e.* that $u_i(x) = \sum_k u_{ik}(x)$.

[15] In principle, some weights may even be negative, even though this seems ethically quite unnatural.

[16] That does not mean that all choices are actually observed, at least in a revealed preference framework. It is the role of the positive analysis to make inferences about hypothetical choices on the basis of actual and observed choices. The key point however is that these inferences are not done on the basis of some metaphysical or ethical considerations about "rational" preferences, but through the identification of choice patterns *given* ancillary conditions.

and objective way. However, as Loewenstein and Haisley (2008, 219) put it, there will "be many years, if ever, before we are able to interpret patterns of brain activation to make inference about what types of choices should count as welfare enhancing". That is, *positive* economics will probably never be of any help to determine which choices or preferences should be taken into account in the welfare analysis. This may well remain an ethical problem. Still, the virtue of the SWFL approach is at least that it makes the planner's value judgments explicit and thus more transparent.

A more serious problem is related to the way the SWFL can deal with the issue of autonomy. The problem lies in the more or less total disappearance of the *person* from the analysis. Consider Bernheim and Rangel's approach once again. As it has been noted in the preceding section, choice-inconsistency across ancillary conditions makes it difficult to Pareto-rank social alternatives either because of the great numbers of weak multi-self Pareto optima or the possible non-existence of strong multi-self Pareto optima. The case where, for each person, selves' preferences agree is unproblematic regarding the consumer sovereignty principle since obviously we fall back to standard welfare analysis. Correspondingly, an ethically plausible reformulation of expression (1) is the "generalized utilitarian" SWFL (Blackorby, Bossert, and Donaldson 2005):

(2) $U(x) = \sum_i g(u_i(x))$, for $i = 1, \ldots, n$ and $g(.)$ any increasing function.

In the special case where $g(.)$ is affine (or if it is the identity transformation of $u(.)$), (2) corresponds to the standard utilitarian SWFL, provided that utilities satisfy full cardinal measurability and comparability. Obviously, (2) is (strongly) Paretian with respect to *persons*: if person i ranks alternative x over alternative y (*i.e.* $u_i(x) > u_i(y)$) while for every other persons j we have $u_j(x) \geq u_j(y)$, then $U(x) > U(y)$. As it is well known, such Paretian SWFL cannot account for non-welfare information because the latter would imply in some cases the violation of the Pareto principle (Blackorby, Bossert, and Donaldson 2005). This is the essence of Sen's early result about the so-called Paretian-liberal Paradox (Sen 1970a). From this point of view, any welfarist framework seems to be ill-adapted to deal with autonomy-related issues (Sen 1979). However, as far as debates about paternalism are concerned, I shall suggest that the alternative selves framework has an even more profound problem.

The SWFL approach corresponding to expression (1) goes further than Bernheim and Rangel's behavioral welfare economics because it makes possible to rank social alternatives even when one or several persons do not make consistent choices across ancillary conditions, *i.e.* when a person's selves do not have consistent preferences. An intriguing feature of (1) is also that it does not satisfy the Pareto principle applied to persons: it is perfectly possible that all members of the population prefer x to y but that (1) ranks y over x. This is due to the fact that a person's utility function is not necessarily additively separable in her selves; additive separability is a feature of the planner's preferences only. Accordingly, this is precisely what paternalism is all about: to go against individuals' will for what is taken to be their own good. Contrary to the laundered preferences approach, the planner makes use of the persons' actual preferences and choices only to evaluate social alternative but he does not give them the same

weight. In this sense, the planner is respectful of the consumer's sovereignty. At the same time however, selves become the locus of the welfare analysis instead of persons.

The implications of the alternative selves framework for the individualistic view of rights have already been noted by Weyl (2009). If one takes the alternative selves framework seriously, then an implication is that the status of the individual or the person as a privileged unit of agency (compared to collectives) is problematic. Modern political philosophy and most law systems consider the person to be the main locus for rights and more generally for deontic powers. The main reason is that the status of the person as a unit of agency has traditionally been considered unproblematic. For example, individual agency implies that persons are – at least up to some limit – responsible for their acts. They also have rights for instance regarding their freedom. However, the figure of the person as a unitary self is currently falling apart. This is not only due to the results of behavioral economics but more generally to many advancements in neurosciences. Philosophers also have started to rethink the received view about the ontology and normative status of the person. Derek Parfit's (1984) reductionist theory of personhood is the most significant attempt to reflect about the moral implications of giving up the standard view of the person. Parfit convincingly argues that there is nothing in the fact of being the same person than the psychological connectedness between a sequence of temporal selves: Adam's present self is in direct continuity with his yesterday self and he is psychologically connected with more remote selves through a sequence of direct psychological relations. Parfit calls this psychological connectedness "relation *R*" and argues that it is (exclusively) constitutive of personal identity. Crucially however, he also argues that relation *R* is not axiologically relevant: from a moral point of view, the psychological connectedness between a sequence of temporal selves does not justify to give a special importance to the person. The relation between two psychologically connected selves is no different than the relation between two unconnected selves from the moral point of view.

Parfit's reductionist account has clearly tremendous moral implications. Regarding distributive justice principles, Parfit argues that the reductionist account of personhood widen their *scope* but reduce their *weight*: while these principles were thought to apply only for the relations *between* lives, the reductionist account indicates that they should also apply to the relations *within* (subdivision of) lives. At the same time, the reductionist account considerably undermines the "separateness argument" put forward by Rawls and others against utilitarianism. Indeed, the reductionist account supports a form of *impersonality* not only between persons but also between selves: if distributive principles have no weight between psychologically connected selves (as it is often thought), then they should have also no weight between unconnected selves. Thus a reductionist view about personal identity seems to lead to what John Broome (1991) calls "complete utilitarianism".

As Parfit recognizes, the reductionist account has also potential implications regarding paternalism and autonomy. Given the axiological irrelevance of relation *R*, to prevent oneself from doing harm to one's future selves is no different than preventing a person *i* from harming a person *j*. Hence, while we may "not believe that we have a general right to prevent people from acting *irrationally*", we may "believe that we have a general right to prevent

people from acting wrongly" (Parfit 1984, 321, emphasis in original). The paternalistic SWFL captured by expression (1) takes full advantage of this possibility: utilities *within* persons are traded exactly the same way they are traded *between* persons. Actually, persons simply disappear from the analysis and only trade-offs between selves are taken into considerations. The key difference between "complete utilitarianism" and the paternalistic SWFL is that the latter does not satisfy a condition known as "anonymity" in social choice theory. While the utilities of some selves are taken to be completely irrelevant for the welfare analysis, others are given a special weight. As noted above, this is fully in accordance with Bernheim and Rangel's suggestion to delete some GCS from the choice data set.

It is interesting to note that this approach is starting to enter into standard cost-benefit analysis, if only implicitly. A revealing and striking example is provided by the recent *Food and Drug*'s (FDA) cost-benefit analysis of the "Graphic Warning Label Regulation" in the United States. In response to this study, a team of economists (Chaloupka et al. 2014) has criticized the FDA's measure of the consumer surplus of the regulation. Chaloupka et al. in particular criticize the FDA's hypothesis that warning labels on cigarettes packaging would reduce consumer surplus because of the loss of satisfaction resulting from reduced consumption. This assumption follows directly from the preference-satisfaction account of welfare: given some stable distribution of preferences regarding cigarettes consumption in the population, any regulation reducing the demand for cigarettes (because of a raise in prices or any other mechanisms) reduces the satisfaction of these preferences. However, Chaloupka et al. claim that this is mistaken because this hypothesis fails to account for the fact that many cigarettes consumers are making "irrational decisions". Given the present discussion, one of their arguments is particularly relevant: Chaloupka et al. urge the regulator to not take into account in the welfare analysis the choices (and preferences) of some (temporal) selves. Indeed, they "refer to this as the 'principle of insufficient reason' approach and argue that the benefits to those who started using tobacco products regularly before 18 years of age and who quit in response to FDA regulatory actions should not have any offset for lost consumer surplus" (Chaloupka et al. 2014, 12).

Chaloupka et al. are thus clearly arguing for the introduction of paternalistic considerations into cost-benefit analysis. Although their proposal could also be interpreted in terms of the laundered preferences approach, I contend that it is more appropriately captured by the alternative selves approach. A key element indeed is that the cost-benefit analysis takes into account preferences that have been actually revealed or elicited. For instance, persons who quit smoking due to FDA actions have indeed revealed a (temporally located) preference for not-smoking. However, while paternalistic, Chaloupka et al.'s proposal does not deal with autonomy-related issues. One reason may be that their proposal is only concerned with welfare *evaluation* but not with collective decision. The alternative selves framework precisely allows making welfare evaluations in accordance with the consumer sovereignty principle. Arguably, respect of autonomy only enters at the collective decision stage. It is not clear however how autonomy issues can be dealt with in this framework given that the person is no part of the analysis.

This seems particularly the case if expression (1) is justified on the basis of Parfit's reductionist account. It would be hard for instance to make sense of Thaler and Sunstein's emphasis on the normative significance of autonomy if their underlying ontology was reductionist in Parfit's sense (Ferey 2011).[17] The key issue is to know how the separability across alternative selves at the analytical level translates at the ontological and moral levels. If additive separability is justified on the basis of Parfit's reductionist account, then discussions about autonomy seem irrelevant despite the superficial reconciliation between the preference-satisfaction view of welfare and the endogenous preferences problem. A closer look at Bernheim and Rangel's framework also indicates that selves cannot be the analytical unit of normative economics because independently of the autonomy problem, the methodological choice of deleting some GCS closely depends on considerations related to the *person*'s attributes when she is choosing.

5. Preferences, Values and Voluntary Consent to Paternalistic Interventions

The alternative selves framework offers a way to accommodate the ethical justification of the preference-satisfaction account of welfare (*i.e.* the consumer sovereignty principle) with the problem of endogenous preferences due to "behavioral failures". But it falls short of providing a completely satisfactory way of introducing paternalism in normative economics because "we are left with no preference-based concept of welfare that applies to the person as a continuing entity" (Sugden 2004, 1017). Since a full discussion of paternalism entails to be able to deal with issues related to autonomy and freedom, the fact that the person is no longer a "continuing entity" is a great difficulty. At least, this is so if we assume that *persons*, not selves, are the locus for values related to autonomy and freedom. In this section, I shall argue that normative economics must distinguish between preferences (which can be attributed to selves) and values (which are necessarily owned by persons). In particular, I will suggest that under some conditions, reasonable persons may agree over a *paternalistic collective choice rule* because they recognize that it is in their interests *as persons* to do so. In essence, this is similar to Rawls' (1971) and Dworkin's [(1972); (2010)] "contractualist" understanding of paternalism.

The distinction between preferences and values has been made in social choice theory. In particular, Kenneth Arrow (1963) introduced the distinction between "interests" and "values", the former corresponding to the individuals' wants and desires while the latter refer to their judgments "all things considered". Similarly, Harsanyi [(1953); (1955)] distinguished between an individual's "subjective" and "ethical" preferences. Both distinctions refer to the ability of a rational individual to make judgments about states of affairs that are not limited to narrow considerations about her own interests. Both Arrow and Harsanyi were interested on how societies may aggregate both kinds of judgments. In the fifth chapter of his classic

[17] Interestingly, in the (short) section entitled "Autonomy and Paternalism", Parfit (1984, 321) does not note the tension between his reductionist account and the very notion of autonomy. He only mentions the "well-known objections" against paternalism, that "[i]t is better if each of us learns from his own mistakes" and "it is harder for others to know that these are mistakes".

Collective Choice and Social Welfare, Sen (1970b) went further and noted that the introduction of values may lead to a conflict: "In the models of collective choice, this conflict is an inescapable one. Individual values are relevant to the exercise in two ways: (a) they affect individual preferences R_i, and (b) they are concerned with the choice of collective choice rules CCR. Values reflected in (a) and (b) could easily conflict" (Sen 1970b, 64-5). Arguably, while preferences may meaningfully be attributed to alternative selves, this is less obvious for values, at least in case (b).

In Sen's terms, a collective choice rule CCR is a method of "going to individual orderings to social preference" (Sen 1970b, 22-3). The paternalistic SWFL (1) described above corresponds to one such CCR since it provides a complete ordering of social alternatives on the basis of the individuals' preferences.[18] The reason why values have to be attributed to persons rather than to selves lays in the fact that choice of a CCR can be seen as a problem of *commitment*. CCR are institutionally designed and depend on more or less formal procedures, norms and values. Though they are by no way unalterable, CCR are relatively stable through times. This stability contrasts with the endogenous nature of preferences that has been discussed in the first section. Of course, the choice of CCR may also depend on unstable preferences but at least for some of them it also reflects the long-lasting influence of values that transcend each person. In other words, many CCR, such as the majority rule in a democratic society, are agreed to because they build on values that everyone recognizes as a member of some community or society. Considering this stability, to choose (or to abide to) a CCR is to commit to a particular mechanism to aggregate preferences with the full knowledge that this mechanism may result to a collective choice that may not correspond to one's preferences at some time in the future or in some context. If the choice of a CCR depends on values, then values cannot be owned by selves because selves by definition cannot commit to anything.

That values enter into the choice of CCR is illustrated by the "non-dictatorship" condition in social choice theory. The ethical attractiveness of this condition (and the reason why most of us will consider it to be a necessary axiom for any CCR) comes from the fact that it reflects an attachment to democratic values. Crucially, this attractiveness is quite independent on the actual preferences of the members of the population and on the resulting collective choice.[19] Autonomy and freedom are other values that may drive the choice of a CCR. They can translate into a variety of properties that the CCR must have; for instance, Sen (1970a) has famously argued that a desirable property of any CCR is to be respectful of an individual's decisiveness over some domain of choice. However, the choice of a CCR may also reflect a trade-off between autonomy and welfare. Suppose that while attaching importance to autonomy and freedom, I also recognize that under some circumstances I am prone to make decisions that I will regret and/or for which I am unable to argue on a reasonable basis (for instance, in a deliberative and public setting) that they will enhance my welfare. In other

[18] Actually, this is a bit imprecise since the utility numbers that serve as inputs in (1) reflect the alternative selves' preferences and not the individuals' preferences.

[19] Using Sen's (1970b, 59-61) distinction, preferences thus correspond most of the time to "non-basic judgments" while values are rather "basic judgments", *i.e.* they are unconditional. For sure, this is not always the case in practice.

words, suppose that I know that with high probabilities I will have preferences that I know I will have a hard time to justify. Now, as a person and while recognizing that my welfare in general depends on the satisfaction of my preferences, I may plausibly agree over a CCR that in some well-defined circumstances will not take into account my preferences in these circumstances. Now, generalize this reasoning to a whole population and we have to conclude that it is perfectly reasonable for persons wanting their preferences to be satisfied *in general* to agree over a paternalistic CCR, *i.e.* a CCR that in some circumstances do not take into account the persons' preferences in these particular circumstances.

As Ferey (2011) points out, Rawls (1971, 249-50) has entertained a very similar view about paternalism. In particular, he argued that the *moral* continuity between the selves of the individual could lead rational people to agree on paternalistic principles under a veil of ignorance: "the principles of paternalism are those that the parties would acknowledge in the original position to protect themselves against the weakness and infirmities of their reason and will in society". The adoption of such principles is "to be guided by the individual's own settled preferences and interests insofar as they are not irrational, or failing a knowledge of these, by the theory of primary goods… paternalistic intervention must be justified by the evident failure or absence of reaction and will; and it must be guided by the principles of justice and what is known about the subject's more permanent aims and preferences, or by the account of primary goods" (Rawls 1971, 249-250). That is, individuals placed under a veil of ignorance may agree to set up institutions (*i.e.* CCR) that protect them against their unreasonable desires and inclinations. At the same time, the device of the original position guarantees that the "parties want to guarantee *the integrity of their person* and their final ends and beliefs whatever they are" (Rawls 1971, 250, my emphasis). Rawls' discussion of paternalism is grounded on his substantive theory of welfare and the concept of "primary goods" but as the quote above indicates, it is also compatible with the laundered preferences approach.

A similar contractualist understanding of "justifiable" paternalism has been developed by Gerald Dworkin [(1972); (2010)]. Dworkin suggests that paternalism is ethically plausible in "Ulysses-type" situations where one readily and reasonably recognizes that it is in his interest to have his hands tied. Importantly, he argues that consent "is important" and appears to be "the only acceptable way of trying to delimit an area of justified paternalism" (Dworkin 1972, 77). Moreover, he argues that paternalism is more plausible regarding matters of facts than regarding matters of values: "We have two distinct types of situation in which a man acts in a non-rational fashion. In one case he attaches incorrect weights to some of his values; in the other he neglects to act in accordance with his actual preferences and desires. Clearly there is a stronger and more persuasive argument for paternalism in the latter situation (Dworkin 1972, 79). Nevertheless, autonomy and freedom are values that paradoxically may command to paternalistic interventions: "I suggest that we would be most likely to consent to paternalism in those instances in which it deserves and enhances for the individual his ability to rationally consider and carry out his own decisions" (Dworkin 1972, 83).

Dworkin's notion of *consent* is clearly essential and is related to the above discussion about the commitment to a (paternalistic) CCR. This strengthens my claim that the alternative selves

framework (and thus the preference-satisfaction view of welfare) is inappropriate to introduce discussion about paternalism into normative economics. For the reasons exposed above, only persons can consent to paternalistic interventions on the basis of an "all-things-considered" judgment. Because the alternative selves framework and welfare economics in general are only considering preferences in the welfare analysis, they necessarily missed out a key feature of the debate over paternalism. In some way, it could be argued that this due to the narrow conception of agency and rationality that is constitutive of economics as a whole. Once again this is a point well captured by Sen's writings. In particular, Sen (2002) argues that rationality is constituted by two related features: maximization and reasoning. Maximization is a technical condition: a rational economic agent is someone whose behavioral pattern can in one way or another be identified to some maximand. As Sen suggests in several writings [(Sen 1993); (Sen 1997)] and as Bhattacharyya et al. (2011) formally demonstrate, it is always possible in a revealed preference framework to do so. Maximization in this sense is empirically vacuous, though theoretically useful. What is substantially distinctive of rationality and thus of agency is *reasoning*:

> "Rationality of choice, in this view, is primarily a matter of basing our choices – explicitly or by implication – on reasoning that we can reflectively *sustain* if we subject them to critical scrutiny" (Sen 2009, 180).

According to Sen, rationality is thus partially a matter of *justification* or *justified reasoning*. The preference-satisfaction account of welfare is only concerned with the maximization aspect of rationality: it is not committed to any specific view of agency beyond choice- or preference-consistency. However, a discussion about paternalism involves more than that: whether to defend paternalistic interventions or reject them, we have to recognize that persons are not merely choice patterns and preference orderings but also value-holders that can reflect upon their preferences and the mechanisms to aggregate them.

6. Conclusion

The main claim in this article has been that normative economics, and particularly welfare economics combined with the preference-satisfaction account of welfare, are currently ill-adapted at the face of the current debates over paternalism. The problem goes well beyond the difficulty raised by endogenous preferences for the preference-satisfaction account. More fundamentally, the difficulty lies in the reductive notion of agency and rationality that underlies welfare economics. In particular, autonomy-related issues are particularly difficult to deal with in such a framework, independently of the well-known inability of welfare economics to account for non-welfare issues.

Neither the laundered preferences nor the alternative selves approaches are able to properly account for the debates over paternalism. My suggestion here has been to take seriously the distinction between preferences and values which, though it has been recognized by social choice theorists for several decades, plays no role in contemporary normative economics. One virtue of this distinction is to reintroduce into the debate over "soft" or "libertarian"

paternalism the key notion of *consent*. Indeed, this is crucial given the ambiguity of libertarian paternalists over the precise nature of "nudges" [(Mongin and Cozic 2015); (Hausman and Welch 2010)]. At least some of them seem to depend on some form of manipulation for their effectiveness. This generates great ethical problems, as recognized by some of their proponents (Sunstein 2014). However, normative economics will be unable to contribute to this essential debate it if does not go beyond the preference-satisfaction account of welfare.

References

Angner, Erik. 2015. *Well-Being and Economics*. SSRN Scholarly Paper ID 2551037. Rochester, NY: Social Science Research Network.

Arrow, Kenneth Joseph. 1963. *Social Choice & Individual Values*. Yale University Press.

Bernheim, B. Douglas, and Antonio Rangel. 2007. "Toward Choice-Theoretic Foundations for Behavioral Welfare Economics." *The American Economic Review* 97(2): 464–70.

———. 2008. Choice-Theoretic Foundations for Behavioral Welfare Economics. In A. Caplin & A. Schotter (Eds.), *The Foundations of Positive and Normative Economics*, Oxford University Press, 155-192.

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

Bhattacharyya, Aditi, Prasanta K. Pattanaik, and Yongsheng Xu. 2011. "Choice, Internal Consistency and Rationality." *Economics and Philosophy* 27(2): 123–49.

Blackorby, Charles, Walter Bossert, and David J. Donaldson. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge University Press.

Broome, John. 1991. *Weighing Goods: Equality, Uncertainty and Time*. Wiley.

Chaloupka, Franck J., Kenneth E. Warner, Daron Acemoglu, Jonathan Gruber, Fritz Laux, Wendy Max, Joseph Newhouse, Thomas Schelling, Jody Sindelar. 2014. "An Evaluation of FDA's Analysis of the Costs and Benefits of the Graphic Warning Label Regulation".

Cowen, Tyler. 1993. "The Scope and Limits of Preference Sovereignty." *Economics and Philosophy* 9(2): 253–69.

Dworkin, Gerald. 1972. "Paternalism." *The Monist* 56(1): 64–84.

———. 2010. "Paternalism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2010.

Elster, Jon. 1985. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge University Press.

Ferey, Samuel. 2011. "Paternalisme libéral et pluralité du moi." *Revue économique* 62(4): 737–50.

Gorman, W. M. 1968. "The Structure of Utility Functions." *The Review of Economic Studies* 35(4): 367–90.

Gul, F. & Pesendorfer, W. 2008. The Case for Mindless Economics. In A. Caplin & A. Schotter (Eds.), *The Foundations of Positive and Normative Economics*, Oxford University Press, p.4-39.

Harsanyi, John C. 1953. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61.

———. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63(4): 309–21.

———. 1977. "Rule Utilitarianism and Decision Theory." *Erkenntnis* 11(1): 25–53.

———. 1996. "Utilities, Preferences, and Substantive Goods." *Social Choice and Welfare* 14(1): 129–45.

Hausman, Daniel M., and Michael S. McPherson. 2006. *Economic Analysis, Moral Philosophy and Public Policy*. 2nd ed. Cambridge University Press.

Hausman, Daniel M., and Brynn Welch. 2010. "Debate: To Nudge or Not to Nudge." *Journal of Political Philosophy* 18(1): 123–36.

Hédoin, Cyril. 2015. "Bargaining in the Mind: The Significance of Multiple Selves Models for Positive and Normative Economics.", working paper, University of Reims Champagne-Ardenne.

Loewenstein, G. & Haisley, E. 2008. The Economist as Therapist: Methodological Ramifications of "Light Paternalism". In A. Caplin & A. Schotter (Eds.), *The Foundations of Positive and Normative Economics*, Oxford University Press, 210-245.

Mill, John Stuart. 1859 [2013]. *On Liberty*. Harper Collins.

Mirrlees, James A. (1982). The economic uses of utilitarianism. In Amartya Kumar Sen & Bernard Arthur Owen Williams (eds.), *Utilitarianism and Beyond*, Cambridge University Press. p.77--81.

Mongin, Philippe, and Mikaël Cozic. 2014. *Rethinking Nudges*. SSRN Scholarly Paper ID 2529910. Rochester, NY: Social Science Research Network.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford University Press.

Pinto-Prades, Jose-Luis, and Jose-Maria Abellan-Perpiñan. 2012. "When Normative and Descriptive Diverge: How to Bridge the Difference." *Social Choice and Welfare* 38(4): 569–84.

Qizilbash, Mozaffar. 2009. *Well-Being, Preference Formation and the Danger of Paternalism*. Papers on Economics and Evolution 2009-18. Philipps University Marburg, Department of Geography.

———. 2012. "Informed Desire and the Ambitions of Libertarian Paternalism." *Social Choice and Welfare* 38(4): 647–58.

Railton, Peter. 1986. "Moral Realism." *The Philosophical Review* 95(2): 163–207.

Rawls, John. 1971. *A Theory of Justice*. Oxford University Press.

Rebonato, Riccardo. 2012. *Taking Liberties: A Critical Examination of Libertarian Paternalism*. Houndmills, Basingstoke, Hampshire ; New York, NY: Palgrave Macmillan.

Ross, Don. 2005. *Economic Theory And Cognitive Science: Microexplanation*. MIT Press.

Samuelson, PA. 1938. “A Note on the Pure Theory of Consumer’s Behaviour.” *Economica* 51(17).

Sen, Amartya Kumar. 1970a. "The Impossibility of a Paretian Liberal." *Journal of Political Economy* 78(1): 152-157.

———. 1970b. *Collective Choice and Social Welfare*. Holden-Day.

———. 1979. “Utilitarianism and Welfarism.” *The Journal of Philosophy* 76(9): 463–89.

———. 1991. *On Ethics and Economics*. Reprint edition. Oxford, UK; New York, NY, USA: Wiley-Blackwell.

———. 1993. “Internal Consistency of Choice.” *Econometrica* 61(3): 495–521.

———. 1997. “Maximization and the Act of Choice.” *Econometrica* 65(4): 745–79.

———. 2002. *Rationality and Freedom*. Harvard University Press.

———. 2009. *The Idea of Justice*. Harvard University Press.

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

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

Sugden, Robert. 1985. “Why Be Consistent? A Critical Analysis of Consistency Requirements in Choice Theory.” *Economica* 52(206): 167.

———. 2004. “The Opportunity Criterion: Consumer Sovereignty without the Assumption of Coherent Preferences.” *The American Economic Review* 94(4): 1014–33.

———. 2008. “Why Incoherent Preferences Do Not Justify Paternalism.” *Constitutional Political Economy* 19 (3): 226–48.

Sunstein, Cass R. 2012. “Storrs Lectures: Behavioral Economics and Paternalism, The.” *Yale Law Journal* 122: 1826.

———. 2014. *The Ethics of Nudging*. SSRN Scholarly Paper ID 2526341. Rochester, NY: Social Science Research Network.

Sunstein, Cass, and Richard Thaler. 2003. *Libertarian Paternalism Is Not An Oxymoron*. SSRN Scholarly Paper ID 405940. Rochester, NY: Social Science Research Network.

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Revised & Expanded. Penguin Books.

Weyl, E. Glen. 2009. “Whose Rights? A Critique of Individual Agency as the Basis of Rights.” *Politics, Philosophy & Economics* 8(2): 139–71.