From Utilitarianism to Paternalism: When Behavioral Economics meets Moral Philosophy

Cyril Hédoin

Dans Revue de philosophie économique 2015/2 (Vol. 16), pages 73 à 106
Éditions Vrin

ISSN 1376-0971
ISBN 9782711652136
DOI 10.3917/rpec.162.0073

Article disponible en ligne à l'adresse
https://www.cairn.info/revue-de-philosophie-economique-2015-2-page-73.htm
from utilitarianism to paternalism: when behavioral economics meets moral philosophy

Cyril Hédoin*

Résumé

Mots-clés : économie comportementale, paternalisme « soft », philosophie morale, personnalités multiples, économie du bien-être

* University of Reims Champagne-Ardenne, REGARDS research center in economics and management. Contact: cyril.hedoin@univ-reims.fr

A previous version of this paper has been presented at the AFEP conference at the University of Bordeaux in July 2013. The author thanks the two anonymous reviewers for their useful comments that have helped to enhance the quality of the article. The author is responsible for all errors and omissions that may remain.

revue de philosophie économique / volume 16, n° 2

© Vrin | Téléchargé le 16/09/2023 sur www.cairn.info (IP: 35.160.27.221)
Abstract
Most behavioral economists take the normative implications of their experimental findings to be broadly paternalistic. They tend to suggest that the results of behavioral economics logically entail the extension of the set of public interventions on the market. In this article, I show that this conclusion follows from an implicit normative reasoning that is unsustainable because behavioral economists remain committed to standard welfare economics. I suggest that the behavioral economists’ defense of paternalism can be understood as an attempt to maximize a social welfare function taking into account the fact that individuals make incoherent choices. But this defense depends on a theory of rational preferences that behavioral economists do not have. Moreover, a defense of paternalism in a welfarist framework leads to downplay the agency dimension of persons. Alternative defenses of soft paternalism may exist but likely require that normative behavioral economics gives up welfarism.

Keywords: behavioral economics, soft paternalism, moral philosophy, multiple selves, welfare economics

Classification JEL : A13, B41, D03, D60

1. INTRODUCTION

Among the small group of recent research programs that has appeared in economics during the last thirty years, behavioral economics is undoubtedly the most debated and the most influential one. Through laboratory experiments essentially, behavioral economists demonstrated that the behavior of human agents reveals systematic deviations from the predictions of standard rational choice theory in economics. Though the external validity of the experimental results (i.e. whether or not results are due to the artificial experimental setup) and their true economic significance are still under debate, it is now widely acknowledged that individuals are not fully rational in the technical sense assumed by mainstream economics. Initially moved by an exclusively descriptive and positive scientific endeavor, behavioral economics has taken recently a normative turn with behavioral economists inquiring into the normative implications of their findings. Since individuals are not rational in a meaningful sense, these behavioral “anomalies” must have consequences in terms of efficiency, wealth and happiness. Richard Thaler and Cass
Sunstein’s defense of “libertarian paternalism” [(Sunstein & Thaler 2003); (Thaler & Sunstein 2009)] and Colin Camerer et al.’s notion of “asymmetric paternalism” (Camerer et al. 2003) are two significant examples of normative statements building from the experimental findings of behavioral economics. Similarly, in a more theoretical perspective, some authors have started to build the foundations for a “behavioral welfare economics” whose purpose is to extend welfare analysis to cases where agents make inconsistent choices [(Bernheim & Rangel 2007); (Bernheim & Rangel 2009)].

Most behavioral economists take the normative implications of their experimental findings to be broadly paternalistic. They tend to suggest that the results of behavioral economics logically entail the extension of the set of public interventions on the market. Traditionally, public intervention has been considered as justified (though not necessarily) in cases of market failures (externalities, public goods). Behavioral economists defending some form or another of paternalism claim that public intervention may be justified even without market failures since we know that people make choices that are suboptimal for themselves: either they regret their choices (in particular when they must make intertemporal decisions) or they do not simply realize that the choices they have made are bad for them.

Various forms of paternalistic policies are proposed by behavioral economists but they mostly fall into what is sometimes called “light” or “soft” paternalism (Loewenstein & Haisley 2008), i.e. policies that are not directly coercive in that they do not restrain the choice set of the choosing individuals nor force them to act in a particular way. Lightly paternalistic policies only steer people’s choices by no coercive means toward what is considered to be the best options for them. Soft paternalism’s “hardest” policies take the form of taxes, such as for instance the “sin taxes” aiming to reduce the consumption of unhealthy products such as alcohol or cigarettes [(O’Donoghue & Rabin 2003); (O’Donoghue & Rabin 2006)]. But softer policies merely consist in modifying the frame under which people make choice, taking advantages of the various “behavioral biases” that make one’s choices a function of the way the decision problem is framed (framing effect, loss aversion, endowment effect, reference-dependent preferences, …). This is the essence of Thaler & Sunstein’s libertarian paternalism and Camerer et al.’s asymmetric paternalism. According to
the proponents of these kinds of paternalism, public policies should be “nudging” or “de-biasing” people’s choices without the recourse of any form of coercion. Broadly, it is argued that it is possible and desirable to help those persons who make mistakes without significantly impeding everyone’s freedom to choose, including the freedom of those who are not subject to behavioral biases.

Such normative extensions of behavioral economics give rise to important difficulties regarding the standard theoretical framework of normative (welfare) economics. Indeed, standard welfare economics has been from its very beginning grounded on the postulate identifying welfare or well-being with the satisfaction of people’s preferences. However, by demonstrating that individuals’ choices are significantly incoherent, behavioral economics casts serious doubts on the relevance of this normative criterion for judging welfare. Basically, the preference-satisfaction criterion states that if an agent prefers the state of affairs where \( x \) obtains over the state of affairs where \( y \) obtains, then she will be better off if \( x \) obtains rather than \( y \). This criterion has generally been given two non-exclusive normative interpretations (McQuillin & Sugden 2012b, p. 555): the well-being interpretation and the freedom interpretation. The former rests on the assumption that individuals generally know and prefer what is good for them. Hence, “it is good for them that their preferences are satisfied” (McQuillin & Sugden 2012b, p. 555). The latter follows directly from the fact of granting freedom of choice (“consumer sovereignty”) a normative value. Given that choices reveal preferences, then preference-satisfaction as a normative criterion follows from freedom of choice.

Combined with the empirical findings of behavioral economics, both interpretations seem to lead to a reformulation of the preference-satisfaction criterion. If persons’ choices fail to reveal coherent preferences, then it is no longer possible to identify their choices with their “true” preferences, thus undermining the foundations of both the well-being and the freedom of choice interpretation.

---

1. Throughout the paper, I define incoherence or inconsistency of choices in the technical sense where individuals’ choices fail to reveal a complete and transitive set of preferences for a sufficiently and significantly large set of choice situations. The causes for such inconsistency may be plural: hyperbolic discounting, framing effect, endowment effect, and so on.

2. Actually, McQuillin & Sugden distinguish a third normative interpretation of the preference-satisfaction criterion, the happiness interpretation. However, it may be better seen as a distinctive welfare criterion. I briefly return on it below.

revue de philosophie économique / volume 16, n° 2
Acknowledging this last point, the purpose of this paper is to make more transparent and to critically evaluate the relationship between the positive results of behavioral economics and the normative claims made by several behavioral economists regarding the implementation of soft paternalistic policies. The point is not only to emphasize the fact that in general the behavioral economists’ defense of paternalism lacks an explicit normative reasoning making clear what are the values and ethical assumptions lurking behind it. More significantly, I argue that the fact that behavioral economists remain committed to standard welfare economics makes such a normative reasoning hardly sustainable. The paper presents two arguments to defend this claim. Therefore, I will proceed in two general steps by first proposing a reconstruction of the normative reasoning that leads behavioral economists to defend paternalistic policies, and then by showing that this normative reasoning is unsuccessful because it suffers from problems inherited from standard welfare economics. More precisely, I argue that the behavioral economists’ defense of soft paternalism lacks an account of rational preferences that rational choice theory and standard welfare economics cannot offer. I also argue that an alternative way to sustain this normative reasoning based on a reductive theory of personhood and a multiple selves framework, though compatible with welfarism, does not work because it downplays the agency dimension of persons. Thus, I suggest that proponents of paternalism cannot avoid to ground their normative extension of behavioral economics on an explicit moral doctrine (i.e. a systematic philosophical account of what is good or right), which here should take the form of a substantive theory of the good, presumably giving up welfarism. Whether or not such a theory is conceivable and relevant is an issue that I left others to take up with 3.

Before advancing, I shall note that my argument has some more or less strong connections with several recent writings on the topic of behavioral economics and paternalism. The rise of paternalism

---

3. My argument should not be understood as an attempt to dismiss all kinds of paternalistic policies. I only intend to claim that even if one accepts the empirical findings of behavioral economics while retaining the standard framework of welfare economics, it does not follow that paternalism is automatically justified. There may be good philosophical and moral reasons to recommend paternalism, but they are generally not addressed by behavioral economists.
resulting from the results of behavioral economics has been addressed by Saint-Paul (2011) in a somewhat dramatic fashion. Saint-Paul urges economists to give up the consequentialism constitutive of normative economics because of the threat “post-utilitarianism” (a modified utilitarianism acknowledging the results of behavioral economics) poses on individual freedom. As a remedy, Saint-Paul argues for the necessity to give to freedom and individual responsibility a lexicographic priority: “individual freedom and responsibility must be recognized as central social values rather than as derived instruments for implementing the calculus of happiness” (Saint-Paul 2011, p. 4). Glen Whitman (2006) expresses concerns similar to Saint-Paul’s but opposes an economic argument to paternalistic policies. Whitman rightly notes that, through the multiple selves framework, the behavioral economists’ argument for paternalism relies on a concept of “internalities” that is formally identical to the externalities concept. However, while they have focused on the Pigouvian solution to internalities (in particular through the implementation of sin taxes), behavioral economists have totally ignored the Coasean solution which consists in letting the agent’s various selves to (intrapersonally) bargain over the internalities. In other words, Whitman suggests that in many cases, paternalistic policies are not needed to address behavioral biases because individuals may voluntarily adopt personal rules successfully managing most of their behavioral inconsistencies. Whitman’s argument is ingenious and is partially supported by psychological and social evidences that people are sometimes able to enforce psychological intrapersonal commitments. But I think that actually it takes the multiple selves framework too literally. Moreover, it is a fact that many persons procrastinate, regret choices made in the past or are unable to honor their commitments. One reason for this is that “intrapersonal contracts” are hard to enforce, meaning that transaction costs are likely to be too high in many cases for the Coasean solution to be workable. The most convincing case against soft paternalism has been made by Robert Sugden in a set of recent writings. Sugden’s arguments range from technical demonstrations that incoherent preferences do not impede the efficiency of the market process under plausible assumptions [(Sugden 2004); (Sugden 2008); (McQuillin & Sugden 2012a)] to a sophisticated critique of the philosophical underpinnings of welfare economics which are kept by
behavioral economists [(Sugden 2011); (McQuillin & Sugden 2012b)]. Regarding the latter, Sugden argues that standard welfare economics and light paternalism share an inadequate approach regarding the viewpoint of the welfare assessment, the “view from nowhere” to use a term coined by Thomas Nagel (1986). Sugden proposes an alternative approach to reconcile behavioral and normative economics, using the normative criterion of opportunity instead of preference satisfaction. Though I am sympathetic to Sugden’s contractarian perspective, my argument will follow a different path – even if ultimately it will reach the same end. Rather than denying the relevance of the “view from nowhere” shared by behavioral and welfare economics, I will argue that even from this viewpoint paternalism does not logically follow from behavioral economics.

Following this introduction, the rest of the article is made of five sections. In the next section, I briefly return on the notion of paternalism and more particularly of light (or soft) paternalism. In the third section, I propose a reconstruction of the normative reasoning of proponents of soft paternalism. I present a formal statement of what I call the “behavioral social welfare function” that the paternalistic and impartial benevolent dictator of welfare economics is deemed to maximize, following the descriptive results of behavioral economics. The fourth and fifth sections single out respectively two key arguments against this normative reasoning on the basis of the preceding reconstruction: one the one hand, behavioral economists lack a theory of rational preferences that neither utility theory nor welfare economics can offer; on the other hand, though a defense of paternalism in a welfarist framework is available through a reductionist theory of personhood, this leads to downplay the agency dimension of persons. The sixth section briefly concludes the article and highlights two ways for normative behavioral economics to account for the inadequacy of its argument for paternalism: the first keeps the informed desire or preference view of welfare but substitutes an objective list theory for the current subjectivist approach; the second gives up the informed desire view as a whole and possibly welfarism.

4. I am not interested here in the issue of whether paternalism is actually respectful of liberal or libertarian principles. Nor will I discuss the nature and the content of the freedom concept that is implicit in the writings of proponents of soft paternalism. For interesting discussions on these two topics, see respectively Grüne-Yanoff (2013) and Ferey (2011).
in general. However it is not clear how behavioral economics can support paternalistic policies in these alternative moral frameworks.

2. THE LOGICAL ARGUMENT FOR SOFT PATERNALISM

I present in this section a reconstruction of the argument developed by behavioral economists to defend soft paternalism. This argument relies on several key definitions and on an important moral premise. A good place to start is Gerald Dworkin’s definition of paternalism (Dworkin 2010): “Paternalism is the interference of a state or an individual with another person, against their will, and defended and motivated by a claim that the person interfered with will be better off or protected from harm”. According to Dworkin, three conditions are necessary and sufficient to consider that $X$ acts paternalistically toward $Y$ by doing (or omitting) $Z$:

1) $Z$ (or its omission) interferes with the liberty/autonomy of $Y$.

2) $X$ do $Z$ without the consent of $Y$.

3) $X$ interferes because $Z$ will improve the welfare of $Y$, including preventing his welfare from diminishing, or in some way promote the interests, values or good of $Y$.

As Dworkin points out, these conditions are far from being free from any ambiguity. Condition 2 has the somewhat odd consequence that one cannot explicitly consent to paternalistic policies. In a democratic society where paternalism takes the form of public policies designed and implemented by elected officials, it has to be weakened in the following way:

2*) $Y$ is unable to do $Z$ by herself and may (tacitly) consent or not that $X$ do $Z$ at her place.

The third condition is the key one. Obviously, it necessitates to distinguishing between the various reasons why someone may want to interfere with another person. Moreover, this condition implies that paternalism rests on a concept of welfare or, more generally, of goodness.

5. Accordingly, it is not necessary that $Z$ actually improves the welfare of $Y$ for $Z$ to be identified as paternalistic. The fact that $X$ believes (even wrongly) that $Z$ will improve the welfare of $Y$ is sufficient.
In short, one cannot justify (or criticize) paternalism without a well-defined concept of the good. At this point, behavioral economists interested in normative matters separate into two groups: a first group of behavioral economists, following the lead of Daniel Kahneman et al. (1997) and Richard Layard (2006), are trying to get “back to Bentham” by defining a substantive criterion of welfare, happiness, measured by what they call “experience utility”. This interpretation of the preference satisfaction welfare criterion clearly departs from the common understanding entertained by most economists, and so I will not have much to say about it. I shall only note that beyond the practical difficulties associated with any measure of “happiness”, the “Back to Bentham” approach rests on a philosophical doctrine (“hedonic utilitarianism”) far from being consensual. The second group keeps with the traditional preference satisfaction criterion of welfare economics, interpreted following either the well-being or the consumer sovereignty interpretations: welfare is defined as the degree to which individuals' preferences are satisfied. This approach is explicitly endorsed by Bernheim & Rangel’s “behavioral welfare economics” [(Bernheim & Rangel 2007); (Bernheim & Rangel 2008); (Bernheim & Rangel 2009)] or by O’Donoghue & Rabin’s discussion of “optimal paternalism” and sin taxes [(O’Donoghue & Rabin 2003); (O’Donoghue & Rabin 2006)]. Though less explicit, this is also the approach endorsed by proponents of “libertarian” or “asymmetric” paternalism. This appears clearly through statements such as the following one: “Drawing on some well-established findings in behavioral economics and cognitive psychology, we emphasize the possibility that in some cases individuals make inferior decisions in terms of their own welfare – decisions that they would change if they had

6. Note moreover that even if one accepts hedonic utilitarianism and takes happiness as the good to be maximized, it remains to be shown that the persons who do not behave according to the axioms of utility theory suffer from a happiness deficit. Evidence is lacking, to say the least (Berg & Gigerenzer, 2010), and it would be very surprising to find any relationship between one’s level of rationality and one’s level of happiness. From this point of view, the critique I develop below regarding the lack of justification to give any normative credence to the axioms of utility theory is also relevant here.

7. Actually, in Bernheim & Rangel’s behavioral welfare economics, the primitive is not preference but choice. This is unproblematic though since in a revealed preference framework such as Berneim & Rangel’s one choices and preferences are synonymous. Their aim is precisely to extend this framework to the cases where agents make “anomalous” choices that may not reflect their true preferences.
complete information, unlimited cognitive abilities, and no lack of self-control” (Sunstein & Thaler 2003, p. 1162, my emphasis).

Sunstein & Thaler obviously identify one’s welfare with one’s well-informed choices, i.e. choices that reveal “rational” or “true” preferences that one should have provided she would not be the subject of behavioral biases [(Qizilbash 2012); (Sugden 2008)]. The idea that it is necessary to distinguish actual choices and preferences from ideal or informed choices and preferences obviously follows from the results of behavioral economics. This is precisely the source of the tension between behavioral and welfare economics I have singled out above. But it should be noted that it is far from being new and that it has been defended by many philosophers under the name of “informed-desire theory”. Even economists firmly opposed to paternalistic policies have pointed out the relevance of distinguishing actual and informed preferences when doing welfare analysis (e.g. Harsanyi 1996).

I shall return below on the idea that welfare should be identified to the degree of satisfaction of “rational” or “informed” preferences since it is the key argument in the defense of soft paternalism by behavioral economists. For the moment, it is sufficient to modify the third condition in Dworkin’s definition of paternalism to accommodate the specific welfare conception of light paternalism:

3*) X interferes through Z because Y’s informed or rational preferences will be more satisfied with Z than without.

A main concern with paternalistic policies is that they are threatening to undermine individual freedom and autonomy. By the way, this is constitutive of Dworkin’s definition of paternalism. Behavioral economists argue for a form of paternalism the most respectful of freedom and autonomy that is possible to have. In particular, with the exception of sin taxes, they recommend policies that make use of minimal or of no coercion. Most of the time, paternalistic policies should operate through the design of the decision problem by taking advantage of the propensity of individuals to be the subject of behavioral bias such as framing effects or reference-dependent preferences. For instance, to increase their employees’ participations in 401(k), employers may simply change the default rule by making enrollment automatic unless employees choose specifically otherwise. Or to prevent consumers from taking impulsive and ill-considered
decisions, the legislator may impose “cooling-off” periods allowing consumers to freely change their mind. In both cases, no coercion is employed (except toward sellers who have the obligation to return the money to consumers who have changed their mind and returned the products), with only a marginal interference with autonomy and freedom. Thus, we reformulate condition 1 in Dworkin’s definition as follow:

1*) \( Z \) interferes minimally with the freedom/autonomy of \( Y \)

Since it is clear that behavioral economists intend to legitimate this specific form of paternalism, it should be normatively understood: a paternalistic policy is good (desirable, legitimate) because it permits to enhance people’s welfare, with the proviso that it does not undermine too significantly individual freedom and autonomy. More precisely, I contend that behavioral economists implicitly make use of what I call the Logical Argument for Soft Paternalism:

*The Logical Argument for Soft Paternalism:*

a) A policy \( P \) is softly paternalistic if it responds to the conditions (1*), (2*) and (3*).

b) If individuals were fully rational and were choosing according to their informed preferences, they would do \( X \).

c) Because individuals are not fully rational and do not choose according to their informed preferences, they actually do \( Y \).

d) If \( P \) is implemented, individuals will do \( X \) and will improve their welfare over the state of affairs where they do \( Y \), while undermining only minimally individual liberty and autonomy [by (a), (b) and (c)].

e) It is morally warranted to interfere such as to improve individuals’ welfare when it undermines only minimally individual liberty and autonomy (*moral premise*).

f) \( P \) is morally warranted [by (d) and (e)].

Obviously, the moral premise is crucial in this logical argument. I will be concerned with the relevance of the concept of “rational” or “informed” preferences constitutive of the definition of welfare entering into it and with the meaning of the notion of autonomy. Sections 4 and 5 argue that this moral premise is hard to defend for

---

8. See Thaler & Sunstein (2009) for numerous other examples.
behavioral economists as long as they remain committed to standard welfare economics.

3. CONSTRUCTING THE BEHAVIORAL SOCIAL WELFARE FUNCTION

Normative economics as a whole can be classified as “welfarist”. Theories of welfare are generally either formal or substantive (Hausman & McPherson 2006, 135). A substantive theory indicates the things or states of affairs that are intrinsically good. A formal theory of welfare specifies how to find what things or states of affairs are good for people, but do not take position on their nature or identity. Hedonic utilitarianism with its emphasis on happiness is a great example of the former. Welfare economics and its preference satisfaction criterion is an instance of the latter. As a formal theory of welfare, welfare economics then states that it is good to satisfy people’s preferences the most.

As this is well-known, Pareto-optimality is the main criterion used in welfare economics to judge the desirability of a state of affairs. A state of affairs \( S \) is optimal or efficient according to the Pareto criterion if it is not possible to increase the welfare of at least one person without reducing the welfare of at least another person. As a corollary, a state \( S' \) is a Pareto-improvement over \( S \) if the welfare of at least one person increases when moving from \( S \) to \( S' \) while other persons are at least as well off. The Pareto criterion is particularly appealing because it is a unanimity criterion. Moreover, it does not rest on the assumption that the welfare of two persons is comparable. Indeed, interpersonal comparisons of utility or of welfare remain controversial within economics. Still, a significant part of normative economics is explicitly grounded on the assumption that interpersonal comparisons of welfare are possible and meaningful. In this case, normative economics is not only welfarist but can also be said to be utilitarian: “Utilitarianism states that “society” should be organized so as to yield the greatest possible level of welfare, where it is assumed that there is some way of comparing and adding welfare across individuals” (Saint-Paul 2011, p. 2, I emphasize). To refer to utilitarianism then “means that some social planner tries to maximize such an aggregate [of the persons’ welfare],
regardless of the specifics of how the aggregate is constructed” (Saint-Paul 2011, p. 26).

As I note above, normative behavioral economics is clearly welfarist. Bernheim & Rangel’s (2007) behavioral welfare economics does not rely on interpersonal comparisons of welfare and uses the Pareto criterion as its sole normative criterion. But some works on soft paternalism are explicitly utilitarian since they rely on interpersonal comparisons of welfare. O’Donoghue & Rabin’s work on optimal paternalism and sin taxes is an example of such a combination of behavioral economics with utilitarianism. However, the question of whether interpersonal comparisons of welfare are warranted is only secondary here. Rather, the main issue is how to determine one’s welfare given the fact that one is subjected to various behavioral anomalies. Since most individuals are prone to make mistakes and to regret their choices, this must mean that at least some of the preferences revealed by one’s actual choices do not enhance one’s welfare. In the preceding section, I noted that most behavioral economics sidestep the problem by identifying welfare with the satisfaction of “rational” or “informed” preferences. Moreover, normative behavioral economics follows standard welfare economics by adopting the “view from nowhere” viewpoint. This is pretty clear in the behavioral economists’ normative writings which regularly speak of a “planner” or a “choice architect” that should make the decision the individuals would make if they had perfect foresight, no weakness of will and more generally if they had chosen according to well-informed preferences. The decision problem of the benevolent dictator of behavioral welfare economics is thus the following: to choose the state of affairs (a social situation or alternative, the frame for a specific decision problem) that satisfies as much as possible people’s rational preferences.

I will now give a precise and formal expression of the choice that a benevolent dictator would make, knowing and taking into account the behavioral biases that affect the members of the relevant population. As it is traditional in normative economics, we assume a purely disinterested planner whose sole aim is to maximize the sum

9. Literally, a planner will make choices on behalf of individuals, while a choice architect will only frame the decision problem such as to steer people’s choices in a given direction. Formally however, there is not much difference: the point is to identify welfare only with choices revealing well-informed preferences.
of the welfare of the members of the population. Since it takes into account the result of behavioral economics, let call him a “behavioral benevolent dictator”. A famous representation theorem demonstrated by John Harsanyi [(Harsanyi 1955); (Harsanyi 1977)] states that if i) the preferences of the \( n \) members of the population satisfy the rationality postulates of expected utility theory, ii) if the preferences of the planner satisfy the rationality postulates of expected utility theory and iii) if at least one person prefers the state of affairs \( x \) over the state of affairs \( y \) and none of the other person prefers \( y \) over \( x \), the planner will prefer \( x \) over \( y \), then the planner must reveal preferences corresponding to the following social welfare function over all states of affairs \( x \):

\[
W(x) = \sum_{i=1}^{n} \alpha_i U_i(x) \quad \text{with} \quad \alpha_i \geq 0 \quad \text{for} \quad i = 1, \ldots, n.
\] (1)

Harsanyi’s aggregation theorem thus states that if the members of the population and the benevolent dictator are rational in the sense of expected utility theory, and if one accepts the Pareto principle, then the preferences of the benevolent dictator can be represented by a social welfare function defined as the weighted sum of the utilities of the \( n \) members of the population. It should be noted that this theorem does not assume interpersonal comparisons of utility; that implies that the weighing coefficients \( \alpha_i \) may vary across the population. If interpersonal comparisons of utility are allowed, then we must have \( \alpha_1 = \ldots = \alpha_n \) with the utility function of each agent rescaled such as to make them comparable. As I argued above, this issue is only secondary here.

If one accepts the empirical results of behavioral economics, then one can hardly assume that the first assumption of the above theorem is satisfied. A significant part of behavioral economics (in particular the writings on the so-called “prospect theory”) is devoted to show that

---

10. Harsanyi argued that his aggregation theorem was a vindication of utilitarianism. Actually, this is not the case because the fact that an agent’s preferences satisfy the axioms of expected utility theory does not make necessary to represent them through an expectational utility function. As it has been widely documented [e.g. (Sen 1986); (Weymark 1991)], if one represents the agent's preferences with a non-expectational utility function, then Harsanyi's theorem no longer holds. As a consequence, what Harsanyi has established is a formal statement to represent the aggregated preferences of the members of a population. Still, the fact that social welfare is additively separable is an ethically significant result, particularly in the present context.
real persons fail to behave as predicted by expected utility theory. In particular, if agents have incoherent preferences, then their behavior cannot be meaningfully represented by a utility function, even less an expectational one. However, following the multiple selves framework, I assume that behavioral inconsistencies across time or across decision frames are due to the coexistence in one person $i$ of several selves $k$, each endowed with coherent preferences satisfying the expected utility axioms. For each person $i$ and each selves $k$, it is thus possible to define a utility function $u_{ik}$. Behavioral inconsistencies follows from the fact that depending on the time one makes a decision, or depending on the framing of the decision problem (e.g. the reference point), or depending on the choice menu, all the actual choices of person $i$ will not necessarily reflect the preferences of the same self $k$. If the selves’ preferences are not the same, then it is possible that the same person makes two incoherent choices under two different frames (which I will assume is the case here). For each person $i$, which self $k$ is active depend on an “ancillary condition”, denoted $d$. Following Berhneim & Rangel (2008, p. 159), I define an ancillary condition as “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”. Ancillary conditions thus capture every characteristics of a decision problem that are beyond the control of the chooser, but that cannot be considered as affecting her welfare. A rational agent with coherent preferences must make coherent decisions across all ancillary conditions $d$; but agents defined by multiples selves and endowed with inconsistent preferences will not make the same choices in two identical decision problems that vary only regarding the ancillary condition.

As an illustration, suppose that an agent $i$ is presented with a decision problem $D$ characterized by a set of alternatives $X$ and an ancillary condition $d$, $D = (X, d)$. Denote $C(D) \subseteq X$ the non-empty set of chosen items in $X$ given $D$. For two ancillary conditions $d'$ and $d''$ and for two alternatives $x, y \in X$, it is quite possible that we observe $C(X, d') = x$ and $C(X, d'') = y$. However, in the multiple selves framework, it will be considered that the two choices

---

11. Ancillary conditions must be therefore distinguished from the “states of nature” in standard decision theory. This definition also implies that the behavioral benevolent dictator cannot choose among a set of ancillary conditions.
\[ C(X, d') = x \text{ and } C(X, d'') = y \] are not made by the same self with the same preferences. As a consequence, despite the fact that the behavior of agent \( i \) is globally incoherent (and thus not amenable to be represented as one form or another of maximization), coherency can be assumed at the level of the selves. I denote \( u_{ik}(x;d) \) the utility function of the self \( k \) of an agent \( i \) when the state of affairs \( x \) obtains, given ancillary condition \( d \) in the background\(^{12}\). The function \( u_{ik} \) is assumed to be expectational and is a representation of the self \( k \)'s preferences. Both preferences and choices are only defined for the set of alternatives \( x \in X \) but not for the ancillary condition. In other words, I assume that one cannot choose the frame under which she makes her choice. I extend this assumption to the benevolent dictator. It is now possible to make use of Harsanyi’s aggregation theorem; we only need to modify the first assumption. Without loss of generality, assume that each agent has two selves \( k' \) and \( k'' \)^{13}, and that each of the \( 2n \) selves’ preferences in the population satisfies the axioms of expected utility theory. Then, by Harsanyi’s aggregation theorem, it follows that the choice of the behavioral benevolent dictator will reveal the following behavioral social welfare function:

\[
W(x) = \sum_{k=1}^{2} \sum_{i=1}^{n} \alpha_{ik} u_{ik}(x;d) \quad \text{with} \quad \alpha_{ik} \geq 0 \quad \text{for} \quad i = 1, \ldots, n \quad \text{and} \quad k = 1, 2. \tag{2}
\]

The behavioral social welfare function corresponding to expression (2) indicates that a planner taking into account the empirical findings of behavioral economics will choose the state of affairs \( x \) that maximizes the weighed sum of the selves’ utilities in the population. As with the original version of Harsanyi’s social welfare function, the value of the coefficients \( \alpha_{ik} \) is undetermined and may vary across the population if the planner does not make “interpersonal” comparisons of utility. In this case, the value the planner chooses to give to each coefficient reflects her value judgment regarding the relative importance of each

---

12. As indicated above, it is important to acknowledge that an agent’s welfare is not a function of the ancillary condition \( d \). To use the important distinction made by Broome (1993), an ancillary condition is a cause of preference but not an object of preference. This explains the semicolon in \( (x;d) \). Since a self \( k \) is activated by an ancillary condition \( d \), to make explicit the ancillary condition is not actually required.

13. For instance, we may imagine that \( k' \) is the present-oriented self and \( k'' \) the future-oriented self in a case where agents reveal intertemporal inconsistencies.
self. But allowing such comparisons would necessitate rescaling the utility functions which would embody similar value judgments.14

As noted above, I have assumed that the planner do not choose under which ancillary condition the agents actually make their choices. But in fact, by choosing the weight she gives to the different selves, she is actually expressing a judgment regarding the rationality of preferences revealed under different ancillary conditions. In Sunstein & Thaler’s (2003) famous cafeteria example, the cafeteria’s director may judge that the customer’s preferences revealed by their choices when the less dietetic items are presented early in the line are not rational or well-informed (i.e. fail to reflect their welfare) while this is the contrary when the most dietetic items are presented first. In this case, the director will choose to give no weight to the preferences expressed in the former case. Indirectly, the director is thus choosing which ancillary conditions are relevant in terms of welfare. Of course, how to determine the weight given to the different selves is the very problem that soft paternalism is facing. In the rest of the paper, I show that behavioral economists and proponents of paternalism do not have an adequate moral doctrine to justify the weighing paternalism is assuming in the behavioral social welfare function.

4. RATIONAL PREFERENCES AND THE WEIGHING OF THE SELVES

The differential weighing of the selves in the behavioral social welfare function necessitates making a value judgment regarding the relevance of the selves’ preferences. The key issue is whether this value judgment can be grounded on a scientific assessment that follows logically from a set of acceptable premises. It is important to note that this issue is not an artifact of my peculiar formalization of soft paternalism in terms of a social welfare function. In their elaboration of a behavioral welfare economics, Berhmeim & Rangel (2007) establishes a normative criterion of individual welfare optima related to the concept of “multi-self Pareto optima”. They note that if for an agent, the same alternative \( x \) is not chosen from the same set of alternatives \( X \) across a set of ancillary conditions \( d \), then a

14. In Harsanyi’s utilitarianism, such value judgments are expressed through what he called “extended preferences”.

© Vrin | Téléchargé le 16/09/2023 sur www.cairn.info (IP: 35.160.27.221)
strict welfare optima does not exist. As they admit, it follows that their normative criterion is not very discerning since the set of (weakly) welfare optima maybe fairly large. They propose a refinement strategy consisting to use nonchoice evidence to discriminate between the selves whose choices are more or less relevant for the planner:

for example, if someone chooses \( x \) from \( X \) under condition \( d' \), where he is likely to be distracted, and chooses \( y \) from \( X \) under condition \( d'' \), where he 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 \((X, d')\) (Bernheim & Rangel 2007, p. 469).

Thus, Bernheim & Rangel are clearly suggesting that the choices of some selves should not be taken into account in the welfare analysis. In my approach in the preceding section, this simply consists to give no weight to these selves in the behavioral social welfare function. My concern here is not the authors’ proposal to use nonchoice evidence to adjudicate between the conflicting selves, even though this is somewhat unorthodox in standard welfare analysis. Rather, what is problematical is the lack of a solid philosophical and logical basis for inferring from the (choice and nonchoice) empirical evidence a set of value judgments setting the unequal weighing of the selves.

Rangel & Bernheim’s suggestion is formally identical to Sunstein & Thaler’s (2003, p. 1162) claim that welfare analysis should only take into account choices that individuals would make “if they had complete information, unlimited cognitive abilities, and no lack of self-control”. This argument relies on a counterfactual reasoning that I have already highlighted in the Logical Argument for Soft Paternalism (claim b): “If individuals were fully rational and were choosing according to their informed preferences, they would do \( X \)”. Surely, one may wonder what is it to be “fully rational” and to have “informed preferences”? Similarly, how can one know (or even make sense of) what choice someone endowed with “unlimited...
cognitive abilities” will make? There is here a subtle difficulty: since welfare is assumed to be reflected only by choices revealing informed preferences, we must be able to determine what are those preferences and to what choices they would lead to. I am not interested here with the epistemic and practical difficulty of determining the relevant method to know the choice an ideally rational agent would make, but rather with the normative significance of associating welfare with the satisfaction of these so-called rational preferences.

To know what a fully rational agent choosing according to informed preferences would do, it is clearly necessary to determine the form and the content of these informed preferences. At this point, normative and positive behavioral economics are deeply intertwined. As Berg & Gigerenzer (2010) document, at its beginning the research program of behavioral economics was devoid of any normative content. Referring in particular to the work of Amos Tversky and Daniel Kahneman during the 1970s and the early 1980s, Berg & Gigerenzer (2010, 18) note that

for some reason, early behavioral economists argued that behavioral economics is purely descriptive and does not in any way threaten the normative or prescriptive authority of the neoclassical model. These authors argued that, when one thinks about how he or she ought to behave, we should all agree that the neoclassical axioms ought to be satisfied.

Even one of the most prominent proponents of soft paternalism asserted twenty years ago that normative inferences should not be drawn from the work of behavioral economists:

A demonstration that human choices often violate the axioms of rationality does not necessarily imply any criticism of the axioms of rational choice as a normative idea. Rather, the research is simply intended to show that for descriptive purposes, alternative models are sometimes necessary (Thaler 1991, p. 138, quoted in Berg & Gigerenzer 2010, p. 18).

As it appears clearly in these quotations, already in its earliest developments, behavioral economics attached a peculiar epistemic value to the axioms of utility theory. However, since at this time the research program did not have a normative agenda, the descriptive results of behavioral economics did not lead to normative conclusions. The behavioral departures from the predictions of utility theory were
interpreted as “anomalies”, to use Thaler’s wording (Thaler 1987), but behavioral economists were not inclined to infer prescriptions in terms of welfare analysis or policies. To characterize behavioral deviations from the axioms of utility theory as “anomalies” or as “mistakes” is of course debatable because it implies to grant to a particular model of human decision an epistemic priority that is far from being self-evident. But as long as normative conclusions were not inferred from this epistemic postulate, it did not weaken behavioral economics as a positive research program.

However, the sustained growth of the literature on soft paternalism during the last ten or fifteen years clearly demonstrate that behavioral economics has now developed a fully normative extension. As the result, what was an epistemic assumption is now becoming a normative commitment that the agents’ behavior *ought* to respect the axioms of utility theory. From a potentially fruitful descriptive theory of the human behavior, utility theory and its concept of rationality have became an exclusive normative theory of what is it to be rational and to act in one’s own interest. Provided that one is willing to make the last step by claiming that it is legitimate to interfere with people’s decisions to help them to pursue their own interest (soft paternalists’ moral premise), then behavioral economics as a positive enterprise transforms into a moral philosophy promoting various forms of paternalism. Once this normative turn is taken, not only we know that real agents do not make consistent choices across time, that their preferences and choices are a function of contextual elements that are secondary regarding the decision problem at stake, that they do not understand probabilities and Bayesian reasoning, and so on. Now, it is claimed that these “mistakes” are detrimental to the agents’ welfare and that they ought to be corrected.

Berg & Gigerenzer (2010, p. 22) sarcastically remark: “It is a great irony that most voices in behavioral economics, purveyors of a self-described opening up of economic analysis to psychology, hang on to the idea of the singular and universal supremacy of rational choice axioms as the proper normative benchmarks against which virtually

---

16. Actually, from its very beginning ordinal and expected utility theory has been a normative enterprise. The very nature of the rationality concept that is constituted by the various axioms of the theory is evaluative and thus potentially prescriptive. However, as long as one assumes that people’s actual behavior at least approximate the axioms of utility theory, no normative inference regarding how people ought to behave are needed.
all forms of behavior are to be measured”. In particular, the axioms of utility theory now become the normative benchmark against which the behavioral benevolent dictator will be able to determine the weight she has to give to the selves in the behavioral social welfare function. I see at least three kinds of arguments to doubt that utility theory and its axioms constitute a relevant benchmark in a welfare analysis: cognitive/logical, evolutionary and factual. The cognitive/logical argument builds on the soft paternalists’ (and more particularly, Thaler and Sustein’s) twofold commitment on a version of the informed-desire view of welfare and on the normative significance of the utility theory axioms. As noted in section 2, Sunstein and Thaler (2003, p. 1162) track welfare on the basis of the preferences that a rational decision maker endowed with “complete information, unlimited cognitive abilities, and no lack of self-control” would reveal. According to Qizilbash (2012, 652), the fact that Thaler and Sunstein explicitly refuse to take position on some objective or substantive list of values on which hypothetical rational decision makers would agree, make their account a close relative to a ideal adviser account such as developed by Railton (1986). According to the latter, an ideal adviser must have “unqualified cognitive and imaginative powers, and full factual and nomological information about his physical and psychological constitution capacities, circumstances, history, and so on” (Railton 1986, p. 173). The benevolent behavioral dictator of the preceding section can then be understood either as the ideal adviser at the population level, or as someone who aggregates the preferences of all ideal advisers in the population.

In the context of behavioral economics, this account faces an obvious difficulty: such an ideal adviser cannot cognitively exist and this leads to great practical difficulties regarding the implementation of paternalistic policies (Qizilbash 2012). This could be partially answered by the claim that this account merely intends to provide approximate guidance for public policies. However, beyond the fact that this has the consequence to significantly downplay the scope and scale of soft paternalistic policies, I shall argue that the soft paternalists’ commitment to the normative significance of utility theory forbids even this half-concession. Indeed, what soft paternalists are assuming is that the ideal adviser must make consistent choices in the sense of utility theory: these choices must be coherent, which implies that they
also must be ‘frame-invariable’, do not be subject of ‘anchoring’ or do result from the application of Bayes’ law. There are strong arguments for being skeptical about the ideal adviser approach of the informed-desire view of welfare; but there are even stronger arguments against this restrictive characterization of the ideal adviser. Actually, there are many instances where a seemingly inconsistent (i.e. not amenable to a representation through a consistent preference ordering or a utility function) behavior seems nevertheless to be reasonable or even ‘rational’ [(Sugden, 1985); (Sen 1997)]\(^{17}\). Basically, to assume that a rational behavior should conform to the axioms of utility theory is arbitrary and does not rely on cognitive or logical foundations.

One may ask whether a definition of rationality as some kind of ‘efficiency’ could not grant utility theory a normative significance. Actually, this is far to be the case. Not only it has been suggested many times that it might be rational (in this ‘efficiency’ meaning) to be irrational (in the ‘consistency’ meaning) (Parfit 1984). This argument is also doubtful from an evolutionary point of view. Presumably, contemporary human agents’ cognitive faculties and behavioral tendencies are the product of their evolutionary history, shaped by the force of natural selection. Even though this does not guarantee that humans’ abilities and capacities are optimal (evolution may be locked in suboptimal pathways), it is certain that they have evolved for their fitness enhancing properties. Several studies suggest that this may be particularly significant in the case of intertemporal “inconsistencies” where hyperbolic discounting has been shown to be an evolutionary efficient cognitive device under uncertainty [(Dasgupta & Maskin 2005); (Robson 2001); (Robson & Samuelson 2007); (Ross et al. 2008)]. Other behavioral “biases” such as the endowment effect or loss aversion may be very likely the result of evolutionary forces that have shaped the cognitive and behavioral dispositions of humans. At a more general level, it can be argued that the use of “fast and frugal heuristics” satisfies a criterion of “ecological rationality” [(Gigerenzer

\[\text{17. I must cite another significant problem which concerns positive economics: it is virtually always possible to describe a decision problem such that a behavioral pattern satisfies some consistency condition. This leads to an underdetermination issue: given a seemingly inconsistent behavior, it is impossible to determine whether the inconsistency results from a mistaken specification of the decision problem or is due to authentic non-rationality. See Bhatatarayna et al. (2011). Obviously, this reinforces the claim that utility theory cannot serve as a normative benchmark for evaluating the adequacy of people’s behavior.}]

**revue de philosophie économique / volume 16, n° 2**
From Utilitarianism to Paternalism

& Selten 2002); (Smith 2009): in a complex and uncertain world, it may be preferable and welfare-enhancing to have a small set of “one-size-fits-all” heuristics than to be programmed to behave systematically according to a coherent set of preferences. It could be argued (as some evolutionary psychologists would be inclined to do) that contemporary humans’ cognitive and behavioral dispositions were quite adapted 10,000 years ago, but that natural selection has not been able to keep with the fast rate of cultural evolution. As a result, these dispositions are now detrimental to their owners. Though plausible, this hypothesis is not factually confirmed. As Berg & Gigerenzer (2010, p. 23) point out, no study have been able to demonstrate that the individuals deviating from the axioms of utility theory suffer any economic losses or live less longer lives. To be more specific, one can seriously doubt that someone behaving assiduously according to the axioms of expected utility theory in a complex and uncertain world as ours would be more successful than someone using, for instance, a case-based mode of reasoning (Gilboa & Schmeidler 2001), as many people do when taking difficult decisions.

Berg and Gigerenzer’s (2010) point about the unjustified normative value of the axioms of utility theory is however only half of the story. Inconsistency of choice is a persistent theme in behavioral economics, both positive and normative. However, it is not so clear that the insistence on consistency is only or essentially due to the fact that behavioral economists give a normative credence to the axioms of utility theory. I would suggest that it is also the result of their commitment toward welfarism and the welfare-as-preference-satisfaction definition of welfare. Let me explain why.

The main objective of welfare economics is to assess and to compare states of affairs in terms of welfare on the basis of some metric or criterion (Pareto criterion, Kaldor-Hicks compensation test, …). This depends on some assumptions regarding how to aggregate individual’s welfare (e.g. whether to allow interpersonal comparisons or not). As soon as welfare is defined as preference-satisfaction, it is necessary to assume that the agents’ preferences are minimally consistent; otherwise, the very notion of welfare would be ill-defined. This is explicitly the problem tackled by Bernheim and Rangel’s behavioral welfare economics: how to preserve a meaningful definition of welfare in spite of the fact that individuals reveal inconsistent preferences?
As I have noted above, their strategy essentially consists in seeing the choices made under different ancillary conditions as if they were made by different agents. The point is that choice or preference consistency is not preserved because of the normative force of the axioms of utility theory. Rather, the fact is that consistency is required for a welfarist framework to make sense at all.

Another point indicates that Berg and Gigerenzer’s critique is only partial. It is clear that the consistency requirement of utility theory is only a formal requirement for rationality. Utility theory (and rational choice theory more generally) does not provide any indication about the preferences that one ought to have. It only states that one’s preferences should be consistent in some well-defined sense. But this is insufficient given the normative project of some behavioral economists. Indeed, many of Thaler and Sunstein’s examples of nudges (for instance the well-known cafeteria and 401(k) examples) suggest that choice inconsistency is not the real problem at stakes. The problem rather lies in the fact that in these particular cases, it seems that people routinely make bad choices for themselves (to eat too much fat food, to not save enough money). Inconsistency is revealed by the twofold fact that sometimes people will regret their past choices (which presumably indicates that their preferences have changed) and that it is actually possible to make them change their behavior simply by changing the decision frame. But then, it is interesting to note that choice inconsistency is no longer the real problem but rather a part of the solution: many nudges discussed by Thaler and Sunstein ostensibly take advantage of the fact that persons make inconsistent choices across ancillary conditions.

This tends to indicate that beyond choice inconsistency, the worry of some behavioral economists is that individuals do not have the “right” preferences. The reasoning in terms of a behavioral social welfare function developed in the preceding section allows to clearly seeing this point: by definition, each self is rational in the sense of utility theory since we have divided a person’s selves such as to make sure that each self has consistent preferences. It follows that the weighing of selves cannot depend on whether or not selves satisfy the axioms of utility theory. Rather, it will depend on the content of the selves’ preferences. Now, of course, a difficulty appears: neither Thaler and Sunstein nor other behavioral economists have an explicit
From Utilitarianism to Paternalism

... substantive theory of rational preferences that could back up the weighing of selves. Intuition or common sense may help here, but the point is that some theory of the good is ultimately required. Scientific results may also give indications regarding what is the right thing to do (e.g. the fact that the regular consumption of fat food increases the risk of fatality due to cardiovascular problems). But this only indicates that positive and normative considerations are deeply entangled. In any case, the definition of welfare as preference-satisfaction retained by behavioral economists is ill-conceived to make the moral premise in the Logical Argument for Soft Paternalism meaningful.

5. DECONSTRUCTING THE INDIVIDUAL: A SELF IS NOT AN AGENT

The preceding section has shown that the positive results of behavioral economics fail to establish a rationale for (soft) paternalism due to the lack of a convincing theory of rational preferences. There is however a second way through which behavioral economics could possibly support paternalistic policies. Once again, the use of a multiple selves framework and the reasoning in terms of a behavioral social welfare function help to make this point clearer: by establishing that people’s make inconsistent choices, behavioral economics shows that a person’s different selves may make choices which are harmful in a way or another to each other. On the ground of a reductionist theory of personhood, this could provide a moral support to paternalistic interventions. This section discusses and rejects this argument.

As I have already suggested, the defense of one form or another of paternalism consists in granting priority to one self over another. In the case of intertemporal inconsistencies, a paternalistic policy will generally grant priority to the future-oriented self over the present-oriented self. In the case of the various forms of the status-quo bias, the self who will be granted priority is the one whose preferences are the more likely to be endorsed by the unbiased individual. Obviously, the multiple selves framework has some important implications regarding the issues of personal identity and of individuality [(Davis 2011); (Ross 2005)]. By dividing the individual agent into several parts, behavioral economists (with the support of modern neurosciences) are putting into question what was a pillar of the Enlightenment.
i.e. the figure of the unitary individual. Some thinkers such as Derek Parfit (1984) have indeed argued that personal identity and the unity of individual “are not what matter”. Beyond mental and physical connectedness and continuity, there is nothing more to the fact of being “the same person”. As Parfit himself suggests, as soon as one endorses such a reductionist account of personal identity, invasive forms of social control seem to be straightforwardly justified. From this perspective, behavioral inconsistencies are even secondary: as soon as one assumes that an individual’s selves are agents in their own rights, and that the duty of society is to protect each agent from transgressions of their “rights”, then interference with an individual’s freedom (assuming that it is meaningful to associate such an adjective to something that is no longer a person) is morally permissible. Behavioral economists promoting soft paternalism only push this logic a step farther, by providing evidence that such transgressions are empirically significant.

The very fact of reasoning in terms of multiple selves is a significant step toward the conclusion that personal identity is no longer relevant in moral matters 18. As Parfit convincingly shows, reductionism regarding personal identity leads to enlarge the scope of morality to the intra-personal realm. However, while a theoretical reasoning grounded on the multiple selves approach may be quite insightful and enhance our understanding of behavioral issues, I shall argue that this is not a normatively appropriate point of view, even if Parfit’s metaphysical reductionism on personal identity is right 19.

Consider the following definition of what an agent is:

---

18. This is not a necessity, though. As long as the multiple selves framework is used as a heuristic device to formalize behavioral biases, it should not lead its user to necessarily endorse paternalistic policies. Some economists have even used the multiple selves approach to defend a conception of the individual as a single being, e.g. Bénabou & Tirole (2002) and the strategies employed in its pursuit. Confidence in one’s abilities generally enhances motivation, making it a valuable asset for individuals with imperfect willpower. This demand for self-serving beliefs (which can also arise from hedonic or signaling motives.

19. Parfit’s arguments are particularly strong regarding temporal selves: there seems to be no deep reason to consider that the connection between a person’s two selves located at two different times is stronger than between two persons living at the same time. However, the multiple selves framework of behavioral economists is not limited to temporal selves. A same person may be meaningfully said to have multiple selves at one time. Here, the case for reductionism is far less convincing.
A is an agent if and only if:

i) **Boundary condition**: A can be relatively easily identified as being A through her agency, including intertemporal agency.

ii) **Narrative condition**: A thinks of herself as a unit of agency and can make sense of the continuity of her decisions made in the past and the decisions she is thinking to make in the future.

The narrative condition implies that an agent views herself as someone “who chooses and lives a particular life […] as things stand, it is *qua* the occupant of this particular body that [she] lives a life, [has] ongoing relationships, realizes ambitions, and carries out plans” (Korsgaard 1989, p. 122). As a unit of agency, an agent is responsible for all her actions and is interested in the consequences not only of her present action but also in the consequences of the future ones. As an agent, I *do* have a reason to care for my future or for whatever may happen to me. Quite sensibly, the narrative condition may be seen as the result of a practical necessity: the unity of agency results from the indisputable fact that an agent’s various selves share the same body. Even split-brain persons must (and generally, succeed in) satisfy the narrative condition by integrating their possibly conflicting conscious neuronal activities into a unique agency (Korsgaard 1989). The boundary condition states that the unity of agency constitutive of personality must be recognized not only by the agent herself, but also by those with whom she interacts. This is particularly important in the case of intertemporal relationships: I am inclined to trust someone making a promise because I recognize her as agent, liable on her past decisions and endowed with a form of continuity in her agency.

On the basis of this definition, I claim that we have solid reasons to accept the following proposition:

*Moral Significance of Agency* (MSA): The scope of morality is restricted to agents. Since selves are not agents, relations between selves are outside the scope of morality.

I will argue that the rejection of proposition MSA entails implausible moral and practical conclusions. First, should we accept the above definition of agency and should we deny to selves the agent label?

---

20. Note that this definition of agency applies not only to human agent but also to some groups of human agents, such as a corporation. Below, I refer to the latter as “groups of agents” even though they are agents in their own rights.
An argument made by Glen Weyl (2009) on the basis of some results in neurosciences indicates a positive answer. If we took seriously the idea of multiple selves, then to grant priority to one of them would be analogous to accept the possibility of dictatorship in something akin to Arrow’s impossibility theorem regarding collective decision making. Dictatorship is generally not considered an acceptable condition in the context of group decision. It is not obvious why we should accept the contrary in the case of intrapersonal bargaining between selves. There is one difference though: in the case of collective decision making, the very notion of dictatorship has a well-defined meaning. Even if it can be ethically or epistemologically argued against, to make a group decision the function of a unique member’s decision is possible and makes sense. Here, the analogy breaks when considering intrapersonal decision making. Indeed, some works in neurosciences tend to suggest that conflicting selves are required for optimal decision making or even for decision making tout court. For instance, Adi Livnat & Nicholas Pippenger (2006) argue that conflicting selves can be understood as the product of evolutionary pressures working at the level of the individual: given physiological and computational limitations, an optimal decision-making system may involve “selfish” selves in conflict with one another. In this sense, dictatorship seems to be an evolutionary non-sense. The multiplicity of selves can then be understood as a functional property required for optimal decision-making. To annihilate (or to weight) some selves seems to run contrary to this functional necessity.

More importantly, as Weyl (2009, p. 146) notes, Livnat & Pippenger’s result has also another possible implication: “We need not see potentially conflicting preferences as separate agents or selves within an individual. Instead […] they might represent independent, distinct criteria for evaluating actions that are not commensurable with one another”. Because an individual’s various and conflicting selves form a coherent whole at a systemic level, it is not clear that one can identify a rational individual with only her “rational” selves. For instance, a striking result in neurosciences is that reason and emotions cannot be meaningfully separated in the decision-making process (Damasio 1994). A rational agent devoid of any emotion is unable to make simple choices because the act of choice implies both reason and emotions. I take this kind of results to imply that, even if
we accept to consider selves as real cognitive or neuronal processes, it is meaningless to take them separately both to explain and to evaluate decision making. In other words, even if one day neurosciences allow us to clearly identify which specific brain region is activated when one is making a particular decision in a particular context (thus somehow satisfying the boundary condition), it would not follow that selves are agents. The functional role of the selves can only be understood when set against the other selves that constitute an agent (the narrative condition). A self cannot reasonably be considered as an agent and thus, according to proposition MSA, moral relations and conceptions should not apply to it.

Acknowledging this definition of what an agent is, proposition MSA seems difficult to reject. Indeed, the argument presented above suggests that contrary to Parfit’s claim, psychological connectedness and continuity are axiologically significant. A rejection of MSA would lead to morally and practically odd conclusions. For instance, assume that selves have deontic powers (a natural implication of the claim that the scope of morality extends to selves). This may lead to the following simple conundrum: an agent’s “emotional-self” $A^E$ causes harm to another agent’s “rational-self” $B^R$, whose rights have not been respected, is asking for reparation. But if we take seriously the idea that deontic powers apply to selves, then $B^R$ should be asking reparation only to $A^E$ but not to $A$’s rational-self, $A^R$. This is simply non-sense: if we accept the claim made above that an individual’s decision is often (or always) the result of the systemic interaction of several selves, one cannot meaningfully attribute deontic powers to a self but not to the others. Moreover, if selves were to be understood as agents endowed with deontic powers, then deontic relationships should apply not only between the selves of a single individual (as it is mostly the case in soft paternalism), but also between the selves of different individuals as in my example. At the time of writing, this is science-fiction. I conclude that the multiple selves framework does not offer a normatively appropriate point of view.

Even if this defense of paternalism was successful, it would not be available to the behavioral economists like Thaler and Sunstein who emphasize the “libertarian” or autonomy dimension of their...
normative position, as Ferey (2011) rightly points out. Freedom and autonomy are values that pertain to persons as agents, not to selves. Any reductionist account of personhood is problematical from this point of view: it is unclear how and why values related to the agency dimension of persons (rather than their welfare dimension) should remain normatively significant if persons are no longer the locus of the normative analysis. Once again, this suggests that behavioral economists create problems for themselves by remaining committed to welfarism. Virtually all the normative discussion in behavioral economics is concerned with the welfare or well-being aspect of the person. Despite occasional (and mostly rhetorical) appeals to freedom or autonomy, the agency aspect is almost never seriously taken into consideration. This is not surprising given the fact that standard welfare economics has essentially ignored this latter aspect. This clearly militates for a broadening of the normative views of behavioral economists, whether or not to defend paternalistic policies.

6. CONCLUSION

Relying on a “common sense moral philosophy”, several behavioral economists defend the implementation of softly paternalistic policies on the ground that individuals significantly fail to be consistent in their choices. I have shown however that this argument for soft paternalism fails because it relies on a dubious moral premise. More exactly, this moral premise depends on an ill-defined concept of welfare which lacks a convincing account of rational preferences. Moreover, I have argued that an alternative defense of paternalism based on a reductionist account of personhood is neither convincing nor available to behavioral economists.

21. Ferey (2011) suggests that John Rawls’ discussion of paternalism in *Theory of Justice* provides more solid moral foundations to libertarian paternalism. In particular, Rawls emphasizes the moral continuity of the person and argues that rational persons may plausibly agree to paternalistic policies behind the veil of ignorance. This is indeed an interesting possibility but it immediately leads to the issue of consent of people. This is a crucial issue given the fact that the effectiveness of many nudges proposed by Thaler and Sunstein seems to depend on some form or another of manipulation. I cannot discuss this point here as it would take us far beyond the problem tackled by this paper.

22. See Sen (1987, chap. ii) for an insightful discussion of the duality between the well-being aspect and the agency aspect of the person.
The argument developed in this paper can be usefully – though a little imprecisely – summarized as the claim that soft paternalism has not solved adequately the reconciliation problem identified by McQuilin & Sugden (2012b). This may lead to two views regarding the way the ‘logical argument for soft paternalism’ can be changed. These views are related but of unequal strength. The weaker view is that normative behavioral economics should abandon its commitment to a “subjectivist” version of the informed-desire view of welfare. Other versions of the informed-desire view rely on a ‘prudential value’ account or an ‘objective theory of reasons’ of the kind developed by Parfit (2011). These versions are not vulnerable to the cognitive argument which greatly limits the normative relevance of the subjectivist version (Qizilbash 2012). But there has been a clear reticence among the proponents of soft paternalism to engage in such a moral and philosophical debate. Moreover, these versions are far from automatically justifying paternalistic policies, whether soft or not. The stronger view is that normative behavioral economics should abandon its commitment to the whole informed-desire view of welfare. But this is not clear how this can be done in the context of welfare economics. Indeed, this would probably entail to give up the preference-satisfaction criterion of welfare economics. A more general approach to welfare, possibly based on some kind of happiness metrics, may eventually lead some results in behavioral economics to normatively justify some paternalistic policies, though this is also doubtful as I noted above (see n. 6). The necessity to take into account the agency aspect of persons highlighted at the end of the preceding section even suggests giving up welfarism as a whole. Still, whatever one’s position on this issue, this clearly indicates that behavioral economics has to seriously engage with moral philosophy.

REFERENCES

104 Cyril Hédoin


© Vrin | Téléchargé le 16/09/2023 sur www.cairn.info (IP: 35.160.27.221)


