Value Commitment, Resolute Choice, and the Normative Foundations of Behavioral Welfare Economics

C. Tyler DesRoches

Arizona State University

Forthcoming in *Journal of Applied Philosophy*

1. Introduction

Joe has arrived for lunch at Carolyn's Cafeteria. Carolyn, the owner of the cafeteria, has displayed several options for her customers. Joe is hungry and only has enough money to purchase either a slice of cake or a piece of fruit. What is the best option for Joe? Many would be inclined to affirm that the healthy option is best for Joe, independent of Joe's subjective attitudes, including his desires or preferences. However, according to revealed preference theory, the canonical theory of microeconomists, whatever Joe prefers is the best option for Joe. On this account, the satisfaction of Joe's revealed preference makes Joe better off, whether any such option is actually healthy or unhealthy.

Behavioral welfare economists have long questioned the satisfaction of revealed preferences as the normative foundation of welfare economics. Given what they have discovered about human psychology over the last several decades, including the endowment effect, the role of attention in decision-making, and the framing effect, most behavioral economists agree that it would be a mistake to accept the satisfaction of revealed preferences as the normative criterion of choice. Experiments have consistently shown that, for a variety of reasons, agents do not possess stable context-independent revealed preferences.

In response, many behavioral economists have suggested that what makes agents better off is not the satisfaction of revealed preferences, but 'latent' or 'true' preferences, which may not always be observed through choice.[1] These preferences can 'float free' of choice.[2] While such authentic or genuine preferences may appear to be an improvement over revealed preferences, some philosophers of economics have argued that they face insurmountable epistemological, normative, and methodological challenges (Infante *et al.* 2016a; 2016b).

This article introduces a new kind of true preference – *values-based preferences* – that blunts these challenges. Agents express values-based preferences when they choose in a manner that is compatible with a consumption plan grounded in a value commitment that is normative, affective, and stable for the agent who has one (Tiberius 2000; 2008; 2018). To have such a plan and to act on it is to adopt a particular kind of strategy when confronted with choice situations. Agents who choose according to their plans are resolute choosers (McClennen 1990; Gauthier 1997). My claim is that while values-based preferences do not apply to every choice situation, this kind of true preference provides a rigorous way for thinking about classic choice situations that have long interested behavioral economists and philosophers of economics, such as 'Joe-in-the-cafeteria.'

The article proceeds as follows. The next two sections consider latent or true preferences and elaborate on the Joe-in-the-cafeteria example. Sections 4, 5, and 6 wrestle with value commitments, values-based preferences, and resolute choice. Section 7 considers challenges to values-based preferences and Section 8 concludes.

2. Joe-in-the-Cafeteria

Richard Thaler and Cass Sunstein (2009, 1-4) begin their book *Nudge* with an example, 'The Cafeteria'. Carolyn, the Director of Food Services for a large city school system, makes decisions that can affect the diets of thousands of children. Carolyn has learned that by changing her food displays, she can increase or decrease the consumption of healthy and unhealthy food items. How should Carolyn arrange her cafeterias? Thaler and Sunstein consider various possibilities, from arranging foods to maximize profits to arranging them at random, but they argue that Carolyn should arrange the food to make the students best off, all things considered. More specifically, Carolyn should use her position to 'nudge' people towards making choices that improve their own welfare as judged by their own subjective standards (Thaler and Sunstein 2008, 5).

While nudges are central to behavioral economics, this article draws no conclusions regarding them, including their efficacy or ethical permissibility (Bovens 2009; Hausman and Welch 2010). Instead, this article focuses on the prior question: what makes some option, rather than another, better for some agent, such as your average Joe? After all, if Carolyn is to follow Thaler and Sunstein's advice in arranging the options so that consumers are made better off as judged by their own standards, then she will need an account of welfare or well-being for Joe, a fallible person whose choices are influenced by various psychological and contextual factors.

The received view, revealed preference theory, states that if Joe chooses some option *x*, when he might have chosen option *y* instead, then option *x* is revealed to be preferred to option *y*. Joe's choices are consistent if they satisfy the weak axiom of revealed preferences, which requires that, if *x* is revealed to be preferred to *y*, then *y* must not be revealed to be preferred to *x*.[3] On this account, the best option for Joe is the one that he chooses.[4]

Behavioral economists have long rejected this view, however. The voluminous empirical evidence contests the thesis that agents always make considered judgments of their own welfare (Dhami 2016). Given what behavioral economists have learned about human psychology over the last several decades, including the endowment effect (Kahneman *et al.* 1990), the focusing illusion (Schkade & Kahneman 1998), and the framing effect (Tversky & Kahneman 1981), most agree that it would be a mistake to accept observed choice as the normative foundation of their discipline. In many cases, agents are systematically affected by factors that have little or no effect on their welfare, interests, or goals. In all such cases, revealed preferences are unstable and context-dependent and, therefore, preferences revealed through choice cannot serve as the normative foundation of behavioral welfare economics.

What then might serve as the welfare criterion for behavioral economics? Most behavioral economists subscribe to the view that the average Joe's 'latent' or 'true' preferences should serve this role (Thaler and Sunstein 2003; Camerer *et al.* 2003). Unlike revealed preferences, true preferences may not be revealed through choice. Sometimes, true preferences have been defined subjunctively – as the preferences that Joe would have acted upon, had he not been subject to the distorting psychological effects discovered by behavioral economists. They have also been defined non-subjunctively, as Joe's 'authentic' or 'genuine' preferences that may not be revealed through choice. Either way, the problem is that Joe's choices can be distorted by psychological factors, such as limited attention, inferior cognitive abilities, or lack of self-control. These factors can operate against the satisfaction of Joe's true preferences. Consequently, behavioral economists have opted to treat such choices as 'errors' that can be corrected with an intervention, such as a nudge or boost (Grüne-Yanoff 2016).

While true preferences might appear to be a *prima facie* significant improvement over revealed preferences, some scholars have raised epistemological, normative and methodological concerns over such preferences, arguing that the average Joe's true preferences should not serve as the normative foundation of behavioral welfare economics (Infante *et al.* 2016a, 2016b). I will consider each of these challenges in Section 6. First, however, I will extend the 'Joe-in-the-cafeteria' example and introduce a new kind of non-subjunctive true preference, values-based preferences, that can respond to these challenges.

3. Setting the Stage

Suppose that behavioral economists are correct to claim that Joe's revealed preference for cake or fruit is unstable and context-dependent. Which option does Joe 'truly' prefer? What is Joe's true preference? Hausman (2016) invites the reader to suppose *ex hypothesi* that the average Joe is 'generally concerned' about his health and appearance, and that he deeply regrets his occasional sugar binges. Hausman interprets health and appearance as mattering more to Joe than the mere pleasure that would be yielded by his consuming unhealthy options. In this example, Joe is supposed to have a true preference for fruit over cake.

Infante *et al.* (2016b) disagree, however. They extend Hausman's example by suggesting that while Joe insists health and appearance matter to him, life's small pleasures, such as eating unhealthy options, matter too. Joe thinks resoluteness is a virtue but he also thinks there is a place in life for acting spontaneously. If we suppose that Joe has a standing resolution to eat healthily, he also must decide whether any particular choice point is an occasion for resoluteness or spontaneity. In other words, Joe must weigh the various considerations at hand and determine his overall preference, which Infante *et al.* claim is to assume Joe has an 'inner rational agent.'

Moreover, this weighing operation will be influenced by contextual cues. For instance, we should expect that Joe's judgement about the relative importance of different dimensions of his life will depend on what Joe happens to be attending to at any particular time and place. When the cake is displayed prominently, after he weighs various considerations, Joe-in-the-cafeteria is more likely to choose the cake when confronted with a choice between cake and fruit. However, Joe-in-front-of-the-mirror, who is attending to his wasteline, may deeply regret having chosen the cake. Infante *et al.* argue persuasively that if Hausman's argument for Joe's true preference is to apply to this case – of Joe-in-the-cafeteria versus Joe-in-front-of-the-mirror – then it must be obvious that Joe-in-the-cafeteria reasons erroneously and that Joe-outside-the-context-of-choice is the author of Joe's true preference.

This article follows the same line of inquiry by extending this example. However, rather than supposing that Joe possesses a perfectly rational inner agent, I will assume that Joe has the 'normal' human capacity to deliberate over his consumption choices in the context of a particular choice *and* outside the context of choice. I will argue that, when confronted with a choice between a healthy and unhealthy option, there *are* ways for Joe to reason erroneously. First, however, something more substantive needs to be said about what it means for Joe to be 'generally concerned' with his own health.

4. Value Commitment & Values-Based Preferences

We are supposing that Joe is 'generally concerned' with consuming in a manner that promotes his own health. Joe is not simply concerned with making *one* healthy choice between a slice of cake and a piece of fruit at a specific time and place. Rather, Joe is concerned with making *many* successive healthy choices over time such that his actions jointly constitute a healthy pattern or

plan of consumption. What does it mean for Joe to be 'generally concerned' with his own health? One way to make this idea clear and distinct is to suggest that Joe is committed to the value of his own health.

Following Valerie Tiberius (2012, 2008, 2005, 2000), I will suppose that a value commitment is normative, affective, and stable for the person who has one. Joe has a pro-attitude towards his own health, which leaves him in a positive affective state. Health, for Joe, is a personal normative commitment that he is motivated to pursue (Calhoun 2009). Joe's value commitment is not a fleeting phenomenon, but is psychologically real, and it possesses diachronic stability across disparate choice contexts. This commitment, for Joe, is different from any mere whim, fancy, or preference, all of which can involve fleeting pro-attitudes. Given this relative stability, Joe does not need to deliberate about the significance of his own health in every choice situation, but he accepts it as a matter of course.[5] The value of his health is a relatively fixed point in Joe's deliberations, and it serves as the basis for planning and actions. Significantly, Joe's value commitment is also justified in the sense that he takes himself to have good reasons to be committed to his own health. Upon reflection, he approves of his pro-attitude towards his health and is aware of no circumstances such that his attitudes would become unstable in response to reflecting on them.

Joe's value commitment serves as a means for him to assess how his life is going. There is something good about his own healthy state, and his value commitment has an authority that guides his specific choices and actions. As will be detailed in Section 5 below, when Joe acts on a consumption plan that is grounded by his value commitment, his actions – his choices over time – express what I term his *values-based preferences*. A values-based preference is a kind of non-subjunctive true or genuine preference grounded by a value commitment, which makes the

preference normative, affective, and relatively stable for the agent who has one. While values-based preferences follow a venerable philosophical tradition that insists on a tight connection between values and preferences, values-based preferences cannot be completely understood in terms of preferences alone (Lewis 1989; Grüne-Yanoff and Hansson 2009; Pérez-Carballo 2018). Whereas preferences are typically construed as subjective comparative evaluations that may generate no reasons for action, a value commitment that grounds a values-based preference is a robust pro-attitude that generates reasons for action (Tiberius 2000, 2008; Raibley 2010; Anderson 1995).[6] As Daniel M. Haybron and Valerie Tiberius (2015, 723) state, "to value something and not merely prefer it is to see it as generating reasons for you—as tending to justify responding in certain ways to it and limiting how you might reasonably respond to it." While values-based preferences may not be revealed through choice, and might be described as a subset of second-order or meta-preferences, they can never be explained as mere subjective comparative evaluations. Invariably, values-based preferences are rooted in something that is psychologically real – a value commitment.

To be clear, my claim is not that satisfying Joe's values-based preferences *constitutes* Joe's well-being or that every value commitment Joe has must bear on his well-being.[7] Given Joe's prudential value commitment to his own health, along with the assumption that Joe is informed of the relevant facts about healthy patterns of consumption (from nutrition science, etc.), then we can reasonably suppose that, other things being equal, Joe's healthy state is positively related to his well-being. Health, for Joe, is a *prima facie* good. For the purpose of this article, it will suffice to presume an evidentiary relation between the satisfaction of Joe's prudential values-based preferences and his well-being, which means that, without making any claims about what

constitutes Joe's well-being, the satisfaction of Joe's values-based preferences is evidence for claiming that Joe is better off (Hausman 2012).

5. Resolute Choice

It is one thing for Joe to report that he is committed to the value of his own health and that he forms the intention to follow through on a healthy pattern of consumption, but it is quite another for him to *act* on this value commitment over time. Joe is confronted with a dynamic choice problem – a series of choices that might not serve his concerns well even though each choice in the series at the time of choice seems perfectly well suited to serving his concerns (Andreou 2016).

While philosophers and others have proposed various ways to resolve dynamic choice problems, this article focuses on one strategy – resolute choice – that was originally developed by Edward F. McClennen (1990) and further refined by David Gauthier (1997).[8] On this account, Joe is confronted with a choice among various consumption *plans*. Every plan specifies a complete implementable sequence of consumption choices over time.[9] On the assumption that Joe is properly informed of the alternatives and exercises good judgement, how is Joe to choose a plan? Joe compares and evaluates the *prospects* or expected outcomes of each plan and judges such outcomes to be either acceptable or unacceptable. In this case, we will assume that acceptable consumption plans are compatible with Joe's value commitment to his own health while unacceptable plans are not.[10] Other things being equal, Joe's commitment picks out acceptable patterns of consumption, healthy patterns, from those that are not (unhealthy patterns). During the deliberative process of selecting a particular plan or set of acceptable plans, Joe evaluates the terminal states of affairs associated with feasible consumption plans and his evaluation is based on

an interval measure defined on these states (a utility function or expected utility function) (Gauthier 1997).

For Joe, acting on an acceptable plan requires adopting a strategy for consuming goods and services over time. McClennen (1990) and Gauthier (1997) identify three distinct strategies available to Joe: (1) *myopic* (2) *sophisticated* and (3) *resolute* choice. Joe chooses myopically when he embarks on a plan but executes another plan that seems best to him at the time of choice, without any consideration as to the likelihood that his choice will continue to seem best to him in the future. By contrast, if Joe manages to look ahead and consider what will be best for him in the future, then he is a sophisticated chooser. For example, he might decide to reject a presently attractive plan, when he predicts that he will come to prefer the prospect of an alternative plan. Finally, Joe might opt to be a resolute chooser. This mode of choice subordinates posterior choices to prior choices. Joe chooses an acceptable plan at the outset, commits to it, and holds to this plan rather than considering later on what would be the most attractive option to choose.

So, which strategy should Joe pursue? Should Joe be myopic, sophisticated, or resolute? The conditions under which resolute choice is rationally feasible has been subject to much debate, but for the purpose of this article, I will set this debate to the side and consider an example of a situation where resolute choice is preferable to sophisticated and myopic choice. McClennen (1990) proposes a condition of 'intrapersonal optimality' on the rationality of resolute choice: agents should choose resolutely in cases where there are benefits to both the present self and the future self that will have to be forgone if one does not act resolutely.

Consider an example – *Joe*. Joe is committed to the value of his own health. He is just about to enter Carolyn's cafeteria and he is also aware that he will have to make a choice between healthy and unhealthy options when he arrives. Given his value commitment, Joe prefers the

healthy option over the unhealthy option, everything considered. However, Joe also knows that, at the time of choice, he will be tempted, and may ultimately choose the unhealthy option. Joe is moments away from entering the cafeteria and he truly wants to choose the healthy option. What should Joe do? Which strategy should he choose?

If Joe is a myopic chooser, then, with sufficient temptation, Joe will abandon his plan to choose the healthy option and, at the time of choice, he will choose the unhealthy option instead. If Joe is a properly sophisticated chooser, then he will realize at the outset that he is likely to change his evaluation of his prospects at the time of choice. With this information, Joe concludes that choosing a plan at the outset and sticking to it – without taking further action – is infeasible for a person like him. Resorting to a sophisticated strategy, Joe proceeds to drink a liter of water before entering the cafeteria, believing that, while drinking the water will cause discomfort, it will also reduce his dehydration, the main cause of his cravings for the unhealthy options. With this strategy in place, Joe is able to follow through on his original plan – to ultimately select the healthy option over the unhealthy one. Alternatively, Joe could be a resolute chooser. Joe chooses his acceptable plan at the outset – a plan that is compatible with his value commitment – and despite his change in evaluation of the prospects when confronted with a choice between healthy and unhealthy options, he resolves to follow through on his original plan.[11] Again, what should Joe do?

Joe should be resolute. Why? One need only compare the consequences of Joe being resolute with those of his being sophisticated. Sophisticated and resolute Joe both choose the healthy option. However, sophisticated Joe faces the cost of drinking a large quantity of water before entering the cafeteria. We might reasonably assume that in evaluating his situation, both before entering the cafeteria and afterwards, Joe prefers going to the cafeteria without the need for consuming a large quantity of water in advance. After all, doing this makes Joe rather

uncomfortable. If that is right, and Joe is resolute, then he considers himself better off than he would be were he sophisticated, since he plans to choose the healthy option without the need to drink water in advance, and he also considers himself better-off at the time of choice because even though he does not seek to realize what he then considers his best prospect, he nevertheless knows that he is making a choice that is not his best prospect. His sophisticated counterpart, on the other hand, would have made the healthy choice, but suffered some discomfort. This example shows that, for Joe, resolute choice is a superior strategy to sophisticated choice. Choosing resolutely is intra-personally optimal. Without choosing in this way, Joe must either suffer physical discomfort or abandon his value commitment.

While it should be clear that value commitments are inessential to resoluteness, Joe is our focus and he has a value commitment to his own health. If he chooses according to a healthy pattern of consumption, then he is a resolute chooser. When Joe acts on his plan, which is grounded in his value commitment, his choices are compatible with his value commitment and express his values-based preferences. Quite simply, if Joe's choices are incompatible with his value commitment, then Joe's choices do not express his values-based preferences.

6. Joe, the Impossible Stoic?

So far, the argument has concerned the rationale and feasibility of Joe attending to values-based preferences in welfare assessments. I have argued that, given Joe's value commitment, resolute choice is a superior to myopic and sophisticated choice. However, this conclusion poses a new problem. If Joe is committed to the value of his own health and chooses resolutely, it might appear that he must *never* waver – that Joe must never choose unhealthy options. Is Joe really an impossible Stoic?

This section considers how Joe might successfully navigate between his values-based and non-values based preferences. There are at least three circumstances under which Joe, who is committed to the value of his own health, is justified in choosing unhealthy options: (1) while remaining committed to his original plan; (2) after abandoning his original plan; and (3) while temporarily deviating from his plan (where competing considerations temporarily outweigh health). Otherwise, making unhealthy choices are, for Joe, errors by his own subjective standards. Given that Joe can justifiably choose some unhealthy options while remaining committed to the value of his own health, the following analysis reveals that behavioral economists and others risk overemphasizing the normative significance of satisfying Joe's true preferences. There are special circumstances when Joe does less well *by his own subjective standards* when his values-based preferences are satisfied.

a. Remaining Committed to One's Plan

Joe's healthy pattern of consumption does not necessarily entail that he must choose the healthy option at every choice point. While Joe's commitment to his own health requires that he chooses a pattern of consumption that is, overall, compatible with his value commitment, this leaves open the possibility that Joe can choose some unhealthy options without deviating from his original plan, which is grounded in his value commitment. Unless Joe's selected pattern of consumption only contains healthy options, which is possible, but unlikely, then Joe's healthy pattern of consumption can contain some unhealthy options. Within Joe's healthy pattern of consumption, Joe will have a 'budget' for making some unhealthy choices.

Suppose, for example, that Joe's consumption plan consists of 100 successive choice points. For every choice point Joe must make a decision between consuming a healthy and unhealthy option. Suppose further that Joe chooses 99 healthy options, and 1 unhealthy option.

Given that Joe chooses the healthy option 99% of the time, it seems reasonable to suppose that Joe's pattern of consumption is healthy. Of course, for every additional unhealthy option that Joe chooses, the more reasonable it becomes to characterize Joe's pattern of consumption as 'unhealthy.' In any case, unless Joe's plan calls for only healthy options, then it can accommodate *some* unhealthy options. Built into Joe's plan is a budget for choosing some quantity of unhealthy options. So long as Joe has not exceeded his budget, Joe can choose unhealthy options and these actions are compatible with his plan and his value commitment that undergirds it.

Whether Joe's pattern of consumption is healthy or unhealthy will depend on any number of facts, including facts about Joe's physiology, genetics, and his level of physical activity. There are bound to be borderline cases that make it difficult to judge whether Joe's pattern of consumption is healthy or unhealthy. It would be striking to learn that, if Joe chooses 49 unhealthy options and 51 healthy options, then his pattern of consumption is healthy, but when he chooses 49 healthy options and 51 unhealthy options, then his pattern of consumption is unhealthy. This suggests that there may be no satisfactory hard and fast rule for deciding when, precisely, Joe's plan has become incompatible with his value commitment in the same way that there is no clear rule for deciding whether Joe is bald. In any case, I will set this issue aside for this purpose of this article. My only claim here is that, given Joe's plan of consumption, he need *not* always choose the healthy option over the unhealthy option. Unless Joe's plan contains no unhealthy choices, then Joe can be a resolute chooser and choose some unhealthy options. Joe enacting his healthy pattern of consumption does not preclude him choosing some unhealthy options.

b. Abandoning One's Plan

Until now, I have supposed *ex hypothesi* that Joe is committed to the value of his own health. But, of course, nothing stops Joe from deliberating once again, after his original deliberation, and

reaching a different conclusion regarding his value commitments and plan of consumption. As Tiberius states, "in a deeply reflective moment, one might decide that one no longer has reason to value something one once valued. This is compatible with thinking that when one has a commitment to the value of some end, ordinary instances of practical reasoning are constrained by that value" (2000, 435). Clearly, Joe can reconsider his value commitments at any time (although if Joe never ceases to reconsider his value commitments, then one might reasonably question his commitment). If Joe deliberates and decides to abandon his value commitment, then it would be mistaken to insist that he must continue consuming according to his original plan. If it turns out that Joe no longer cares about pursuing his healthy pattern of consumption, or even feels hostile towards the possibility of choosing healthy options, then it would be a weakness of resolute choice if it required that Joe must stick to his original healthy pattern of consumption no matter what.

c. Temporarily Deviating from One's Plan

There is third way that Resolute Joe is justified in choosing unhealthy options while remaining committed to the value of his own health. Suppose that, for every choice point Joe confronts, he chooses the healthy option because he sticks to his plan. However, while Joe is committed to the value of his own health, he also values spontaneity. When Joe is out on the town celebrating with friends or attending a wedding, he occasionally chooses unhealthy options. One way to make sense of this is to suggest that the preferences Joe occasionally manifests in the presence of unhealthy options are local departures from his normal plan of consumption. If Joe occasionally chooses the unhealthy option when his plan calls for no unhealthy choices, then Joe does not act resolutely. However, Joe cannot be said to completely abandon his value commitment and plan either. Instead, Joe temporarily blocks the application of his plan in a context of particular choice situations.

How are we to make sense of this case? If Joe has a standing resolution to consume the healthy option at every choice point, but he also, on occasion, strongly prefers the unhealthy option, must Joe invariably act on his original plan? At any given choice point, how is Joe to decide whether it is an occasion for resoluteness or spontaneity? To answer this question, I will adapt Gauthier's (1997, 20-21) analysis and apply it to the example of Joe-in-the-cafeteria.

Let us first distinguish between Joe's values-based preferences that he acknowledges when choice is not imminent from his *proximate preferences* that Joe sometimes acknowledges when he is confronted with a choice between healthy and unhealthy options. At any given time, Joe may want to act on his proximate preferences (choose the unhealthy option), but he does not want to act at other times on what would then be his proximate preferences, because they are in conflict with his (all-things-considered) values-based preferences. Joe understands that if, given his proximate preferences, he chooses the action that best realizes his immediate concerns, then he is deliberating in a way that is incompatible with his plan, which is the best realization of his overall concerns, as viewed at *that time* or *at any other time*.

Suppose, again, that Joe's plan consists of 100 choice points. At each choice point, he must choose between a slice of cake (unhealthy option) and a piece of fruit (healthy option). Suppose that when a particular choice is imminent, Joe prefers a slice of cake this time but a piece of fruit on all other occasions. When no choice is imminent, Joe prefers fruit on every occasion. Suppose also that when choice is not imminent Joe prefers 100 pieces of fruit to 100 slices of cake. If Joe chooses based on his proximate preferences, then taking all of his choices into account, Joe will choose slices of cake on each of the 100 occasions. This mode of choice does not best realize Joe's concerns as viewed *at any time* because were Joe to choose on the basis of his values-based

preferences, Joe would choose 100 pieces of fruit, and whether or not a choice is imminent, Joe prefers 100 pieces of fruit to 100 slices of cake.

This example presumes that Joe has two bases of deliberation: Joe's proximate preferences at the time of choice and Joe's preferences at a time when no choice is imminent. If Joe deliberates on the basis of his proximate preferences, then he does less well in realizing his preferences, as they are *at any time*. Joe would have done better if he had deliberated on the basis of his values-based preferences. If Joe evaluates the prospects that would be realized from consistent choices based on his proximate preferences and compares the prospects to those that would be realized from consistent choices based on his values-based preferences, Joe would conclude that the former plan is less favorable than the latter plan. If Joe deliberates on the basis of his proximate preferences, he will do less well in realizing his preferences overall, as they are at any time, than he would have done if he deliberated on the basis of his values-based preferences removed from the context of choice. Joe would be irrational to deliberate on the basis of his proximate preferences.

So far, I have supposed that Joe always evaluates the prospects of choices consistently based on his values-based preferences more favorably. But, what if Joe finds himself in the thralls of a particular choice situation, and there is some unhealthy option available to him that overwhelms everything else, including Joe's ordinary evaluation of prospects? In this case, we can suppose that Joe judges that the benefits of satisfying his proximate preference outweighs the costs of sacrificing his values-based preference.

Joe can recognize in the strength of his proximate preferences what Gauthier (1997, 21) refers to as a *threshold of immediacy*. A proximate preference below this threshold is such that at the time of choice between some healthy and unhealthy options, Joe would choose not to act on

this preference, if it entailed that Joe must therefore act on all proximate preferences of equal strength. A proximate preference above the threshold is one that, at the time of choice, Joe would choose to act on all proximate preferences of equal strength.

The main point here is that if Joe deliberates on the basis of his proximate preferences below his threshold, he will do less well in realizing his values-based preferences, not only as he views them in the context of a particular choice situation, but as he views them at any time, than were Joe to deliberate on the basis of preferences removed from the time of choice, and proximate preferences above the threshold. It would be irrational for Joe to deliberate on the basis of his proximate preference when they are insufficiently strong to cross Joe's threshold of immediacy. However, by the same reasoning, it would be rational for Joe to deliberate based on his proximate preferences (to act against his value commitment) when they are sufficiently strong to cross the threshold.

7. Values-Based Preferences and the Methodological, Normative, and Epistemological Challenges

As mentioned in Section 2, true preferences have been criticized on epistemological, normative and methodological grounds. What are these challenges and how do values-based preferences stand up to them? The epistemological challenge states that, even if people had true preferences, it remains unclear how behavioral economists could know about them. After all, unlike revealed preferences, true preferences may not be revealed through choice. The epistemological challenge is independent of the account of values-based preferences developed in this article. While the burden of inferring values-based preferences falls on the shoulders of behavioral economists who wish to design interventions that nudge or boost people towards the satisfaction of such preferences, that project is a separate issue. With that being said, if Joe in-the-cafeteria has a true

preference for a healthy pattern of consumption that is difficult to infer, this state of affairs only makes it all the more urgent for behavioral economists to focus on developing reliable methods for inferring them. On the assumption that some people, such as Joe, really do have values-based preferences, then the epistemological challenge should be overcome, not used as a reason to claim that people do not or cannot have such preferences. While values-based preferences are more difficult to infer (and measure) than revealed or behavioral preferences (because choices can float free of values-based preferences), this observation alone is hardly a reason to ignore them.

Nor is the empirical basis of latent or true preferences always controversial. Infante *et al.* (2016a) acknowledge that, in some retail markets, where competing suppliers offer the same product priced according to different tariffs, it seems reasonable to assume consumers have a true preference for paying less rather than more, even when consumers choose the more expensive option. In cases like this, representing choices as mistakes or errors defined relative to true preferences for low prices is a reasonable assumption to make when modelling individual choice. In this kind of case, one might insist that the options have an objective ranking, in inverse order of their prices that is independent of the consumer's subjective judgements.

Another potential strategy for inferring true preferences, suggested by Hausman (2016), would involve conducting experiments. In the case of choosing between fruit and cake, one might place both items in equally prominent positions, and advertise an equal amount of nutritional information for each option. Under these settings, if most agents without any impairments to deliberation choose the cake over the fruit, then this would be evidence for claiming that cake is the option latently preferred by of most agents.[12]

Most promising perhaps are the relatively new methods, developed by behavioral economists, for inferring true preferences when choices are unobservable and likely to reflect

errors (Benjamin *et al.* 2014). While the details of these methods are beyond the scope of this article, they have been used to infer preferences that would have been observed from well-informed and deliberated choice data. As such, these methods go a long way to blunt the epistemological challenge, and there is no good reason to think that some variation of this methodology could not be used to infer values-based preferences, specifically.

The normative challenge is a worry about whether some agent, such as a government or policy-maker, is ever justified to nudge others towards the satisfaction of their true preference. The central concern here is that some such interventions may fail to respect people and their actual choices. As Hausman describes, it seems that Infante *et al.* "object to a normative economics that imposes the economist's or policy-maker's judgment of what is good for people rather than simply furthering their choices – or that they object to a normative economics that does this while pretending to conform to people's preferences" (Hausman 2016, 29). First, there is no *pretending* to conform to people's values-based preferences. Either individuals, such as Joe-in-the-cafeteria, have values-based preferences or they do not. Second, it should be clear that the main objective of this article has been to defend an account of true preferences that can serve as the criterion of choice for a class of cases initially popularized by Thaler and Sunstein (2008) and further developed by others (Infante *et al.* 2016a, 2016b; Hausman 2016). Beyond this objective, many questions remain, of course, including those regarding the permissibility and efficaciousness of interventions designed to promote any preferences. While such questions are central to behavioral economics, they are peripheral to the argument defended in this article.

It is worth recognizing that the normative challenge is *not* a special problem for true preferences, including values-based preferences. Furthering the revealed preferences or choices of individuals might be problematic as well. Consider an intervention designed by behavioral

economists to benefit some individuals by furthering their choices, which may not promote their own well-being. It remains an open question whether helping someone satisfy a revealed preference for an endless supply of addictive drugs could ever be justified. Other things being equal, helping to satisfy this problematic preference would seem to be bad for the individual. Promoting the satisfaction of preferences, whether revealed or true, is never a value-neutral decision. Invariably, interventions that promote the satisfaction of some preferences over others face some version of the normative challenge.

Finally, the methodological challenge states there is no psychological foundation for ascribing agents with stable and context-independent true preferences. Responding to this challenge requires either arguing that there is no need for such a foundation, or that there may be an adequate psychological foundation for true preferences. Following the latter pathway, Hausman (2016) has argued that your average Joe-in-the-cafeteria has the capacity for context-independent reasoning, and that he himself can affirm or dispute the existence of his true preferences. But Infante *et al.* reject this line of reasoning by insisting on the context-dependent nature of revealed preferences and claiming that there appears to be no scientifically defensible way to identify features of human psychology that represent true objectives and distinguish them from features that can be classified as reasoning flaws.

While it has been widely documented by behavioral economists that revealed preferences are context-dependent and unstable, values-based preferences are not. As discussed in Section 4, the value commitments that ground values-based preferences are psychologically real, which makes them a potential candidate for serving as the basis for true preferences (Raibley 2010 and Tiberius 2000; 2008). Moreover, they are relatively stable because they are reflectively endorsed (reflective equilibrium). The agent who possesses values-based preferences is committed to a value

that stabilizes these preferences across disparate contexts of choice, even if there are circumstances when the agent fails to act on their true preference. It is critical to recognize that to claim values-based preferences are stable across relevant contexts is not to claim that such preferences, and the value commitments that undergird them, never change (McKerlie 2007). As described in the previous section, agents can always reevaluate their value commitments. However, possessing a genuine value commitment entails that one's commitments are a relatively fixed point in deliberations about how to act or what to choose.

8. Conclusion

Values-based preferences redirect one's attention from individual consumption choices to whole patterns of consumption. What matters for a person such as Joe, who is committed to the value of his own health, is not that he makes one or two healthy choices, but that he chooses over time in a manner that expresses his value commitment, which is normative, affective, and stable.

The foregoing analysis reveals significant lessons for behavioral economists and philosophers of economics. From the purview of values-based preferences, behavioral economists risk overemphasizing the normative significance of true preferences. There are circumstances when agents would do less well – by their own subjective standards – to satisfy their values-based preference than to satisfy some particularly strong (proximate) preference that the agent possesses in a specific choice situation. This result complicates the project of designing and deploying interventions, such as nudges and boosts. After all, even if such interventions were ethical and efficacious, and behavioral economists had knowledge of Joe's values-based preferences, they would still have to decide whether, for any given choice point, they should intervene to ensure Joe

chooses the healthy option (Joe's value commitment is compatible with choosing some unhealthy options).

For philosophers of economics it should be clear that the three – epistemological, normative and methodological – challenges do not paralyze latent or true preferences. On the contrary, this article has argued that one kind of true preferences, values-based preferences, go a long way to blunt these challenges.

Finally, one might object to values-based preferences by arguing that they only apply to a narrow range of choice situations and, therefore, are of limited value. Indeed, there should be no doubt that these preferences only apply to choice situations characterized by an agent, such as Joe, with a value commitment. One might describe such cases as self-acknowledged self-control problems (Sugden 2017; Sunstein 2018; Sugden 2018b). While values-based preferences may not apply to every case, they do provide a rigorous framework for thinking about a class of choice situations that is ubiquitous to behavioral economics and, therefore, ought to be taken seriously.

References

Andreou, Chrisoula. 2016. "Dynamic Choice," *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2017/entries/dynamic-choice/

Angner, Erik. 2018. "What Preferences Really Are," *Philosophy of Science* 85(4): 660–681.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball and Alex Rees-Jones. 2014. "Can Marginal Rates of Substitution be inferred from Happiness Data? Evidence from Residency Choices." *American Economic Review*. 104 (11): 3498-3528.

Binmore, Ken. 1994. *Playing Fair*. Cambridge, MA. MIT Press.

Bovens, Luc. 2009. "The Ethics of *Nudge*." In Grüne-Yanoff, Till and Sven O. Hansson (Eds.) *Preference Change: Approaches from Philosophy, Economics and Psychology*. Dordrecht: Springer, 207-219.

Calhoun, Cheshire. 2009. "What Good is Commitment?" *Ethics* 119 (4): 613-641.

Camerer, Colin, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue, and Matthew Rabin. 2003. "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism.'" *University of Pennsylvania Law Review* 151: 1211-1254.

Dhami, Sanjit. 2016. *The Foundations of Behavioral Economic Analysis*. Oxford: Oxford University Press.

Gauthier, David. 1997. "Resolute Choice and Rational Deliberation: A Critique and a Defense." *Noûs* 31 (1): 1-25.

Grüne-Yanoff, Till. 2016. Nudge versus Boost: How Coherent are Policy and Theory? *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science, 26* (1-2), 149-183.

Grüne-Yanoff, Till and Sven O. Hansson (Eds.). 2009. *Preference Change: Approaches from Philosophy, Economics and Psychology*. Dordrecht: Springer.

Haybron, Daniel M. and Valerie Tiberius. 2015. "Well-Being Policy: What Standard of Well-Being? *Journal of the American Philosophical Association*. 1 (4): 712-733.

Hausman, Daniel M. 2016. "On the Econ within." *Journal of Economic Methodology* 23 (1): 26-32.

Hausman, Daniel M. 2012. *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.

Hausman, Daniel M. 1992. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

Hausman, Daniel M. and Brynn Welch. 2010. "Debate: To Nudge or Not to Nudge." *The Journal of Political Philosophy*, 18 (1): 123-136.

Heathwood, Chris. 2019. "Which Desires are Relevant to Well-Being?" *Noûs* 53 (3): 644-688.

Infante, Gerardo, Guilhem Lecouteux, Robert Sugden. 2016a. "Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioral Welfare Economics." *Journal of Economic Methodology* 23 (1): 1-25.

Infante, Gerardo, Guilhem Lecouteux, Robert Sugden. 2016b. "'On the Econ within': A Reply to Daniel Hausman." *Journal of Economic Methodology* 23 (1): 33-37.

Kahneman, Daniel, Jack L. Knetsch and Richard H. Thaler. 1990. "Experimental tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* 98: 1325-1348.

Lewis, David. 1989. "Dispositional Theories of Value." *Proceedings of the Aristotelian society, Supplementary Volume* 63 113-137.

McClennen, Edward F. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge: Cambridge University Press.

McKerlie, Dennis. 2007. "Rational Choice, Changes is Values over Time, and Well-Being." *Utilitas* 19 (1): 51-72.

Pérez-Carballo, Alejandro. 2018. "Rationality and Second-Order Preferences." *Noûs* 52 (1): 196-215.

Raibley, Jason R. 2010. "Well-Being and the Priority of Values." *Social Theory and Practice* 36 (4): 593-620.

Schkade, David A. and Daniel Kahneman. 1998. "Does Living in California Make People Happy? A Focusing Illusion in Judgements of Life Satisfaction." *Psychological Science* 9: 340-346.

Sen, Amartya. 1971. "Choice Functions and Revealed Preference." *Review of Economic Studies* 38: 307-17.

Sen, Amartya. 1973. "Behaviour and the Concept of Preference." *Economica* 40: 241-59.

Sen, Amartya. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy & Public Affairs* 6 (4): 317-344.

Stanovich, Keith E. 2008. "Higher-Order Preferences and the Master Rationality Motive." *Thinking & Reasoning* 14 (1): 111-127.

Sugden, Robert. 2004. "The Opportunity Criterion: Consumer Sovereignty without the Assumption of Coherent Preferences." *American Economic Review* 94: 1014-1033.

Sugden, Robert. 2017. "Do People Really Want to be Nudged towards Healthy Lifestyles?" *International Review of Economics* 64: 113-123.

Sugden, Robert. 2018a. *The Community of Advantage: A Behavioral Economist's Defence of the Market*. Oxford: Oxford University Press.

Sugden, Robert. 2018b. "'Better off, as Judged by Themselves:' A Comment on Evaluating Nudges." *International Review of Economics* 65: 9-13.

Sunstein, Cass R. and Richard H. Thaler. 2003. "Libertarian Paternalism is not an Oxymoron." *The University of Chicago Law Review* 70: 1159-1202.

Sunstein, Cass R. 2018. "'Better off, as Judged by Themselves:' A Comment on Evaluating Nudges." *International Review of Economics* 65: 1-8.

Thaler, Richard H. and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Tiberius, Valerie. 2000. "Humean Heroism: Value Commitments and the Source of Normativity." *Pacific Philosophical Quarterly* 81: 426-446.

Tiberius, Valerie. 2005. "Value Commitments and the Balanced Life." *Utilitas* 17 (1): 24-45.

Tiberius, Valerie. 2008. *The Reflective Life: Living Wisely Within Our Limits*. Oxford: Oxford University Press.

Tiberius, Valerie. 2018. *Well-Being as Value-Fulfillment: How We Can Help Each Other to Live Well*. Oxford: Oxford University Press.

Tversky, Amos and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211 (4481): 453-458.

Endnotes

[1] Robert Sugden (2018) argues that preferences should be abandoned altogether and supplanted with the enlargement of options. For the formal treatment of this possibility, see Sugden (2004).

[2] Chris Heathwood (2019) makes a similar claim regarding "desires in the genuine-attraction sense" (as opposed to behavioral desires).

[3] If the weak axiom of revealed preference theory is satisfied, then one can construct a complete, transitive, and continuous revealed-preference ordering from them (Hausman 1992, 19).

[4] For an alternative account of revealed preference theory, see Ken Binmore (1994). For a critique of revealed preference theory, see Amartya Sen (1971; 1973).

[5] As described below, it is consistent with valuing health that Joe will pay no attention to health in many decisions in which the alternatives do not appear to him (at least at the moment) to have different implications for health.

[6] Hausman (2012) has argued that, among economists, preferences are total subjective comparative evaluations. For a critique of this view, see Erik Angner (2018).

[7] This article makes no claim regarding the correct philosophical theory of well-being, such as hedonism, objective list theories, or preference satisfactionism. For substantive values-based theories of well-being, see Jason Raibley (2010) and Valerie Tiberius (2018). The relevant value commitments for this article are *prudential* or self-interested value commitments. This focus does not entail all value commitments are merely prudential, however. Joe might be committed securing the well-being of future generations that will only come into existence after his death, thereby negating the possibility that his commitment yields personal benefit (Sen 1977). While such non-prudential value commitments are both reasonable and ubiquitous, they are also tangential to the purpose at hand, which is to analyze value commitments that yield prudential benefit.

[8] My exposition of this account will be informal. For a formal version, the reader should see McClennen (1990).

[9] For each choice point, a plan specifies a unique choice among those available given the choice it specified at the preceding choice point, if any, and each possible combination of intervening chance events, if any, since the preceding choice point (Gauthier 1997, 2).

[10] It should be clear that some plan A might be acceptable to Joe, while plan b is not, even though plan b is better with respect to health, if it is worse in other regards.

[11] Following McClennen and Gauthier, I am supposing that the resolute chooser does not need *any* costly strategy for resisting temptation, etc. If she did need such a strategy, then she would no longer be a resolute chooser. She would be a sophisticated chooser. Sophisticated choice and resolute choice are mutually exclusive.

[12] However, as Infante *et al.* (2016b) recognize, there remain significant practical challenges to inferring latent preferences and there is reason to believe that such experiments would run into significant problems of their own.