Original Research Article

Politics, Philosophy & Economics
1–17

DOI: 10.1177/1470594X241239987
journals.sagepub.com/home/ppe

Behavioral economics and the evidential defense of welfare economics

Garth Heutel
Georgia State University and National Bureau of Economic Research, USA

Abstract

Hausman and McPherson provide an evidential defense of welfare economics, arguing that preferences are not constitutive of welfare but nevertheless provide the best evidence for what promotes welfare. Behavioral economics identifies several ways in which some people's preferences exhibit anomalies that are incoherent or inconsistent with rational choice theory. I argue that the existence of these behavioral anomalies calls into question the evidential defense of welfare economics. The evidential defense does not justify preference purification, or eliminating behavioral anomalies before conducting welfare analysis. But without doing so, the evidential defense yields implausible welfare implications. I discuss how the evidential defense could be modified to accommodate behavioral anomalies.

Keywords

welfare economics, behavioral economics, preference purification, evidential defense, revealed preference

Introduction

Much of standard neoclassical welfare economics is based on the idea that preferences can be revealed by choices and that those preferences are a guide to evaluating welfare or well-being. This view of the relationship between preferences and welfare is problematic, since there are reasons to be skeptical of a preference-satisfaction theory of welfare

Corresponding author:
Garth Heutel, Georgia State University and National Bureau of Economic Research, Atlanta, GA, USA.
Email: gheutel@gsu.edu

that claims that whatever satisfies preferences by definition promotes well-being. Thus, the philosophical justification for standard welfare economics is on shaky ground. Hausman and McPherson (2009) present an *evidential defense* of the standard methodology of conducting welfare economics by using preferences.[1] According to their argument, rather than committing to the preference-satisfaction theory of welfare, economists can use people's preferences as good evidence of their well-being so long as they are self-interested and well-informed.

This is however false, considering the findings from behavioral economics. Even self-interested and well-informed preferences exhibit "behavioral anomalies," where what is preferred may differ from what contributes to well-being. The existence of these behavioral anomalies presents problems for the standard revealed preference methodology of welfare economics: if the preferences themselves are inconsistent or irrational, how can they constitute welfare or be used in welfare analysis? Thus arises the methodology of behavioral welfare economics called "preference purification": preferences are purified of their behavioral anomalies before they can be used to assess welfare. The evidential defense of welfare economics runs into problems when choices themselves do not give good evidence of welfare. When individuals are subject to behavioral anomalies that create a systematic wedge between their choices and their purified welfare function, then claiming that preferences can guide welfare analysis inappropriately confounds choices with welfare.

The purpose of this paper is to argue that the existence of behavioral anomalies calls into question the evidential defense of welfare economics. My main argument (presented in detail in section "The evidential defense in light of behavioral economics") is as follows. If one wants to continue to apply the evidential defense of welfare economics even in the presence of behavioral anomalies, then one must either first purify these preferences of their behavioral anomalies, or not purify them and use the potentially inconsistent and irrational preferences as the best evidence of welfare. In either case, we encounter difficulties. If preferences are not purified of their behavioral anomalies, then the evidential defense ends up endorsing inconsistent and implausible welfare evaluations. This can be avoided if preferences are purified, but the evidential defense provides no justification for such a purification absent some theory of what well-being is, which the evidential defense purports to be agnostic about. Hausman and McPherson (2009) maintain that the justification is provided by platitudes, though there are no platitudes addressing behavioral anomalies like default bias. There is a substantial difference between correcting preferences from non-self-interest or mistaken beliefs and correcting them from behavioral anomalies.

What one needs instead of platitudes is not a full-fledged theory of well-being but rather claims about what kinds of actions and preferences *do not* provide consistent evidence for well-being. In the final section, I discuss a strategy that could allow the evidential defense to be used to justify welfare economics even in the presence of behavioral anomalies. I propose slightly modifying the evidential defense by dropping the assertion that it need not rely on any theory of well-being. Rather, what is needed is a theory of welfare *errors*, which I argue can be arrived at more easily than a full-fledged theory of welfare, salvaging aspects of the evidential defense.

In the following section, I begin by discussing the standard view of welfare economics typically practiced by welfare economists and in cost-benefit analysis, and I summarize

the evidential defense of welfare economics. I also briefly review the field of behavioral economics and discuss its implications for welfare economics. The section "The evidential defense in light of behavioral economics" presents my main argument, which is that the evidential defense encounters problems in dealing with behavioral anomalies. I conclude (the section "Modifying the evidential defense to accommodate behavioral anomalies") with a brief sketch of a proposal for how the evidential defense could be modified to attempt to address these concerns.

Welfare economics, the evidential defense, behavioral economics, and preference purification

Economists are often concerned about the effects of policy or of markets on people's welfare or well-being. There are two steps to the standard approach. First, they look at people's choices—in particular at their choices in markets—to "tease out" or estimate their preferences. This is the "revealed" part of "revealed preferences."[2] This leads to the second step in welfare economics, which is to equilibrate preference ranking to welfare ranking: if i prefers X to Y then X yields higher welfare for i than does Y. One interpretation of this method is that welfare economists are equating preference satisfaction with well-being. This is the preference-satisfaction theory of well-being.

Many doubt the justifications of measuring welfare in this manner. However, Hausman and McPherson (2009) provide a defense. They defend traditional welfare economics on the grounds that, even though preference satisfaction is not *constitutive of* well-being, it is nonetheless good *evidence for* what promotes well-being. Therefore, conducting traditional welfare economics, like cost-benefit analyses, and treating preferences as if they were constitutive of welfare is a defensible, practical strategy for conducting welfare analysis. I will call their argument the "evidential defense of welfare economics," or just the "evidential defense." The power of this defense lies in its agnosticism about a theory of what constitutes well-being. It emphatically does not equate well-being with preference satisfaction.

Hausman and McPherson (2009) acknowledge that the evidential defense does not do away with all the standard objections to the preference-satisfaction theory of well-being. How does a welfare economist determine what types of preferences are evidence for well-being? Hausman and McPherson (2009) do not provide a strict set of rules, but they emphatically claim that the determination does *not* require a philosophical theory of well-being. Instead, "the way out of the difficulty lies in platitudes concerning well-being rather than through embracing some other philosophical theory" (p. 18). Economists should essentially use their best judgment as to what preferences provide evidence for well-being.

Moving on to a brief overview of behavioral economics, roughly speaking, it is the field of economics that points out that the way that some humans actually behave is often not the way that economics models have traditionally described humans as behaving. A precise and universally-accepted definition of behavioral economics is hard to come by. Angner and Loewenstein (2012) call it "the attempt to increase the explanatory and predictive power of economic theory by providing it with more

psychologically plausible foundations." Mullainathan and Thaler (2000) offer "the combination of psychology and economics that investigates what happens in markets in which some of the agents display human limitations and complications." It is a collection of observed "behavioral anomalies" or "behavioral biases" (I will use these terms interchangeably), which in some way are examples of how people are behaving irrationally.

But what exactly constitutes a behavioral anomaly? They can be defined based on a violation of some consistency axiom (Anand, 1987; Mahmoud, 2017). For example, one might declare that preferences that do not exhibit transitivity violate such an axiom and are irrational. Observing a set of preferences that are intransitive would thus give an example of a behavioral anomaly. These anomalies are often related to an analogous set of preferences that do not violate the consistency axioms, as I explain below. Researchers have identified several such anomalies and have developed several descriptive theories of behavior to accommodate those anomalies. This is the understanding of behavioral anomalies that I will use in this paper, and in particular will use one specific anomaly—default bias—as an illustrative example.

Default bias, or status-quo bias, occurs when the default option disproportionally affects an individual's preferences. One's preference for a health insurance plan (Johnson et al., 2013) or a retirement savings plan (Choi et al., 2004) varies with whatever default option is assigned by one's employer. This pattern of choice is inconsistent with rational choice theory and in violation of consistency axioms. Intransivity can occur if the defaults are offered in a certain order (Lahiri, 2019).

The notion of there being a behavioral anomaly does not imply that it must also be a behavioral mistake or a behavioral error. In this paper, I will reserve the term "mistake" to refer to someone being wrong about some objective fact of the world (what Hausman and McPherson call "mistaken beliefs"). Behavioral anomalies are not mistakes nor based on mistakes. However, behavioral anomalies may or may not be thought of as errors. By "errors" I will take to mean something that is wrong in a welfare-relevant way. That is, if preferences violate some consistency axiom, that makes this an example of a behavioral anomaly. If those anomalous, inconsistent preferences lead the person to make decisions that are not promoting their welfare, then it would be an error. My main argument below in the next section will hinge on the relationship between the definitions of behavioral anomalies and behavioral errors.

What are the implications of behavioral economics for welfare economics?[3] If people's behavior cannot be modeled by rational choice theory, then the link between observed choice and well-being must be called into question. Much of the literature devoted to behavioral welfare economics has taken the approach that has been called "preference purification" (Hausman, 2012; Infante et al., 2016). Preferences determine how people behave and the choices they make, but they do not necessarily determine or reveal well-being. Instead, only purified preferences give us information about well-being. Under this interpretation, behavioral anomalies are errors, and the preferences on which we act deviate from maximizing our well-being.

There are several different ways in which one can enact preference purification. For example, some studies have posited that there are two different utility functions. One utility function describes the way that people behave, while the second utility function represents well-being. Behavioral anomalies are places where those two utility

functions diverge.[4] This is not the only methodology of preference purification available. The practice of preference purification itself has been called into question, for example by Infante et al. (2016) and Sugden (2018). A debate has emerged about the validity of Sugden's criticism of preference purification, and his preferred alternative (relying on the "opportunity criterion"). For example, Thoma (2021) and Bernheim (2021) argue that an alternative form of preference purification based on Bernheim and Rangel (2009) does not suffer from Sugden's critiques, while Dold and Rizzo (2021) argue for an approach based more closely on the work of James Buchanan. The argument that I will present below will apply equally well to various versions of conducting behavioral welfare economics, whether purifying preferences or not.

The evidential defense in light of behavioral economics

I now present my main argument. I consider how one would attempt to use the evidential defense to justify the use of revealed preference welfare economics when people exhibit behavioral anomalies. That is, assume that it is a fact of the world that people's (at least some people's) preferences are inconsistent and predictably irrational, in accordance with various models from behavioral economics. Can I still use the evidential defense to justify conducting welfare economics? There are two broad strategies to take when applying the evidential defense here: either you could purify preferences or not. Not purifying preferences amounts to not treating the behavioral anomalies as anything like an error or anything that the welfare economist would have to correct for. Purifying preferences means treating the behavioral anomalies as errors and correcting for them in some way. Under either strategy of using the evidential defense in the context of behavioral economics, the defense runs into problems, as I will now argue.[5]

The originators of the evidential defense argue that, regardless of the existence of any behavioral anomalies, there is still a requirement that preferences be "purified" in some sense (though perhaps they would not use the term "preference purification"). Sarch's (2015) summary of the evidential defense identifies two conditions necessary for preferences to be good evidence of well-being: that preferences be *self-interested* and that the person is *well-informed*. But there is an important difference between purifying mistaken or non-self-interested preferences and purifying preferences that lead to a behavioral anomaly. Preferences being *mistaken* is distinct from preferences being *biased* in the sense of creating behavioral anomalies. It is worth clarifying this distinction since it will be important to the argument here. Consider the example of default-biased preferences, where the default option affects preferences. There is nothing mistaken in this behavior—it isn't that the person is wrong about the default or what options are available. Rather, the preferences are well-informed but generate the behavioral anomaly of default bias. It is uncontroversial to purify the first type of preferences. Purification of the second type is less straightforward, and it is that purification that I will focus on in my main argument in this paper.

Consider the first strategy of using the evidential defense of welfare economics in the context of behavioral anomalies, which is to *not* attempt to purify the preferences of their behavioral anomalies, so long as those preferences are well-informed and self-interested. By way of example consider default-biased preferences again. If my default retirement

option is plan A, I choose plan A, but if the default is plan B, I choose plan B, despite the fact that the which plan is the default has no effect on either plan's costs or benefits. Each of these preferences is self-interested—they are about me choosing a retirement plan for my own benefit (ignore benefits to spouses or children). And each of these preferences is well-informed—nowhere am I mistaken about what my options are or the costs and benefits of any of my options. If using the evidential defense without purifying preferences of their default bias, one is claiming that the best evidence for what maximizes my well-being when plan A is my default is plan A, but the best evidence for it when plan B is my default is plan B. This seems very unsatisfactory. Preferences may be inconsistent in this way, but it does not seem plausible that *what constitutes well-being* is similarly inconsistent. It cannot be that, evaluated under one default, what promotes my best interests or well-being is a substantially different course of action than what promotes my best interests or well-being under a different default, when the only difference is the option that happens to be offered as a default.

Thus, applying the evidential defense of welfare economics in the presence of behavioral anomalies without attempting to purify preferences of those anomalies leads to some quite unsatisfying and implausible conclusions. The welfare economist ends up being just as inconsistent, irrational, and anomalous as ordinary people are. It does not seem that the defenders of the evidential defense would endorse this interpretation of it. Indeed, Hausman and McPherson (2009) admit that welfare economists must correct for "mistakes, biases, and non-self-interested motives."

We are now left with the second strategy that welfare economists can take when attempting to use the evidential defense in the face of behavioral anomalies: they can attempt to purify the preferences of those anomalies. To be clearer about this purification process, I assert that the evidential defense of welfare economics could be slightly re-stated as such: preferences are the best evidence for what maximizes people's well-being, so long as those preferences are: (1) self-interested, (2) well-informed/not mistaken, and (3) unaffected by any behavioral anomalies that cause people's actions to deviate from what makes them better off.

Admittedly, this third condition is somewhat tautological and therefore somewhat unappealing. *Of course* preferences have to be such that they don't interfere with people making choices to increase their well-being before they can be used as evidence for what constitutes people's well-being. This raises two questions: what types of behavioral anomalies cause people's actions to deviate from what is in their best interests? And, how exactly are preferences that exhibit these anomalies to be purified?[6]

For the first question, the obvious candidates for behavioral anomalies that need to be purified from preferences are the usual suspects: the behavioral anomalies that have been identified by behavioral economists over the past several decades, including default bias. However, the justification for purifying these types of preferences is not immediately clear. A key feature of the evidential defense is that it is agnostic about what constitutes welfare.[7] But, how can we justify purifying preferences of their behavioral anomalies without relying on some theory of welfare? To reiterate my earlier point, this same objection does not apply to the less controversial claim that preferences need to be purified of their mistaken beliefs and non-self-interestedness. I do not need a theory of welfare to claim that preferences based on mistaken beliefs are not good evidence of what

constitutes welfare, nor do I need a theory of welfare to claim that preferences that are not self-interested are not good evidence of what constitutes (individual) welfare. But, it is difficult to see how I can justify claiming that preferences that are well-informed and not mistaken but exhibit behavioral anomalies are *not* good evidence for what constitutes welfare, without a theory of well-being.

Why? To clarify, there are three types of preferences that are candidates for being purified before being used in welfare analysis: mistaken preferences, non-self-interested preferences, and preferences exhibiting behavioral anomalies. According to the evidential defense, the first two types of preferences can be purified without relying on a theory of well-being, and this is correct. If preferences are based on mistaken beliefs, that is, facts about the world that are objectively wrong, then the justification for purification is clear. For non-self-interested preferences, the justification is slightly less obvious but still apparent. Hausman and McPherson give the example of preferences over the existence of endangered species. Someone can prefer that these species continue to exist, and some part of their existence contributes to that person's welfare, because she may get pleasure out of watching nature documentaries of them in the wild. However, "it is hard to see what other contribution to individual welfare the continued existence of these species could make, because it is hard to see how their existence bears on other intrinsic goods" (p. 18–19). This is a plausible justification, based on platitudes and not on a theory of well-being, that non-self-interested preferences ought not to be counted as evidence for welfare. I do not see a similar plausible justification for omitting preferences based on behavioral anomalies from what counts as evidence for welfare, without having a theory of well-being that rules them out.

One might disagree with my claim that there are no platitudes regarding behavioral anomalies that can be used to purify preferences of them. To see this, consider a different example of a behavioral anomaly: time inconsistency leading to present-biased preferences. When people exhibit present bias, they act as if placing extra value on the present period relative to all other periods.[8] This behavioral anomaly can explain (under one interpretation) the act of procrastination. For the same reason that decisions made under default bias do not seem to be reliable indicators of what promotes well-being, decisions made under procrastination do not seem to be reliable indicators of what promotes well-being. This intuition can be described as a platitude regarding when to purify preferences. One might argue, following the argument from Hausman and McPherson on non-self-interested platitudes cited in the previous paragraph, that a suitable platitude can justify purifying preferences of time inconsistency. Thus, we have a platitude justifying the purification of preferences from time inconsistency that parallels the platitude justifying the purification of non-self-interested preferences, according to this argument.

I have two responses to this argument that platitudes also exist justifying the purification of preferences of behavioral anomalies like time inconsistency. First, the example above seems to me to be stretching the definition of "platitude" (admittedly never clearly defined) by relying on rules that are less clear and uncontroversial, therefore less justifiable to use without invoking a theory of well-being. This is the key difference between platitudes about the first two categories of preferences to be purified—mistaken and non-self-interested preferences—and the third category—those exhibiting behavioral

anomalies. It is uncontroversial that we cannot rely on mistaken preferences to guide us towards evaluating well-being. It is also uncontroversial that we cannot rely on non-self-interested preferences, based on the fact (or "platitude") that features of the world that are not related to an individual do not affect that individual's well-being. Extending the argument to the example of procrastination demands more to be justified—preferences featuring procrastination are about the individual, and they are not mistaken. To argue that we have a platitude, analogous to the one that rules out non-self-interested preferences, is implicitly relying on a notion or theory of well-being that is absent from the platitudes governing mistaken or non-self-interested preferences. The platitude described in the previous paragraph implicitly relies on a notion or a theory of well-being, for example, the notion that what contributes to one's well-being must be consistent with a plan that one makes ahead of time rather than consistent with how one deviates from that plan. This is much less clear and less uncontroversial than the platitudes covering the other two types of purifications.

I admit that this first response relies on a rather poorly-defined fine line of demarcation between platitudes that do not rely on a theory of well-being and those that do implicitly rely on such a theory. One might reject this demarcation and believe instead that the purification of preferences featuring procrastination or time inconsistency is also justified based on reasonably uncontroversial platitudes that are agnostic about a theory of well-being. Very well. This brings me to my second response to the argument that platitudes also exist justifying the purification of preferences of behavioral anomalies. While platitudes may exist for some behavioral anomalies, like time inconsistency, they are generally unavailable for most behavioral anomalies. Again, consider the example of default bias. Any platitude that would justify purifying preferences of their default bias would have to rely on some theory of well-being that asserts that well-being cannot depend on default, and thus it would not be agnostic about a theory of well-being. For a platitude concerning default bias, the claim that such a platitude is agnostic about a theory of well-being is even more of a stretch than is the claim about a platitude concerning time inconsistency. As I stated at the beginning of this paragraph, one might argue that purifying preferences of procrastination is justified based on uncontroversial platitudes. I disagree, and furthermore I do not think that one could plausibly make the same argument for purifying preferences of default bias. The procrastination platitude, if one supports the interpretation that it is indeed a suitable platitude not relying on a theory of well-being, is fundamentally about one's preferences being *erroneous*—for example, when you prefer to skip going to the gym you are making an error about what is in your own best interest. One could possibly argue that this claim about preferences being erroneous (not mistaken about objective facts of the world, but erroneous about what outcomes will best contribute to one's well-being) is justified without a theory of well-being because there exist platitudes about welfare errors arising from procrastination due to features of human psychology that we have been aware of for centuries.[9] The discovery of default bias is more recent and so platitudes defending their purification are less likely to exist. Finally, even if one admits platitudes defending purifying preferences of default bias, the method of purifying those preferences cannot be supplied with only platitudes, as I will describe below.

So, it is hard to justify purifying preferences of their behavioral anomalies before using them as evidence of well-being without having some theory of well-being that justifies the purification.[10] Let us now move on to the second question that their proposed purification begs: how exactly would we purify these preferences? As with the first question, here we will run into the same issue: it is difficult to come up with a method for purifying the preferences without relying on a theory of welfare that tells us how.

Again, consider default bias. It is hard to justify a specific way to purify preferences in this case. The closest thing to a standard methodology of preference purification is to assume that the default bias arises from loss aversion and prospect theory and assert that the decision utility function is characterized by prospect theory but that the purified utility function is characterized by expected utility theory (Bleichrodt et al., 2001). It does not seem plausible that this level of specificity in what does and does not constitute evidence for well-being can be justified based on platitudes or on anything short of a theory of well-being. I can think of no platitudes that exist that would justify or explain that the best evidence for well-being consists of preferences that fail to exhibit a kink in the value function at a reference point or probability weighting (two features of prospect theory). Similarly, for present bias caused by time-inconsistent preferences, the standard way of purifying preferences is a rather technical assertion about using one particular discount factor rather than another (O'Donoghue and Rabin, 2006). Now, platitudes can justify the claim that when people have self-control problems, or when they procrastinate, or when they are lazy, that their preferences might not improve their well-being. But it is asking too much of these platitudes to dictate exactly how we can filter out the parts of people's preferences that contribute to well-being from the parts that do not.

One might be concerned that my main argument in this section faces a fundamental dilemma. On the one hand, if behavioral anomalies present a problem for the evidential defense in that they must be purified because they cannot be reliable guides to well-being and no platitudes exist that support their purification, then it follows that there is not enough evidence to conclude that they are welfare errors in the first place. On the other hand, if behavioral anomalies are so clearly irrelevant to well-being, then there is no controversy or problem with purifying preferences of those anomalies; that is, platitudes must exist. Thus, my argument criticizing the evidential defense will either fail the same way the evidential defense fails, or else the evidential defense will face no problems.

In fact there is no dilemma in the argument. The way out is to delineate between instances where preference satisfaction cannot reliably be used to promote well-being, and platitudes that justify purifying preferences of behavioral anomalies. These two groups of things are not identical to each other, though they may appear to be. Uncontroversial welfare errors include those emerging from default bias. It cannot be conducive to my well-being to choose plan A over plan B when plan A is the default, but then conducive to my well-being to choose plan B over plan A when plan B is the default. That is the claim that this is a welfare error. It is a different claim to say that an uncontroversial platitude exists, absent a theory of well-being, that justifies purifying preferences of default bias before conducting welfare analysis. Claims about the existence of welfare errors are different than platitudes justifying preference purification because the former can exist without a reliance on a theory of well-being, though the latter (as

I have argued) cannot. The "leap" from the appeal to platitudes to purify preferences from non-self-interestedness and from mistaken beliefs to purifying preferences of behavioral anomalies is a leap that the original evidential defense glosses over, and the extension is by no means merely a "technicality" or a corollary.

Allow me to expand on this distinction. Consider again the specific example of default bias in retirement plan choice. This is unreliable evidence for well-being, since the default option under which a welfare analysis is conducted (i.e. which preferences the welfare economist is examining) has no bearing on actual welfare outcomes or evidence of such outcomes. These things are clearly unrelated to well-being and so evidence that relies on them or is subject to them is unreliable. Does it follow that we have platitudes that justify purifying preferences of their default bias, in the same way that we have platitudes that justify purifying preferences of their mistaken beliefs? No, not without a theory of well-being. The distinction is that the existence of platitudes implies that there is some way of constructing or defining well-being that implicitly requires a theory of well-being, though such a theory is not required of merely claiming that the preferences provide unreliable evidence for well-being. The "leap" is in going from merely observing that these preferences lead to incompatible welfare conclusions, to claiming we have a way to rid these preferences of what leads to that incompatibility. This "leap" is not present for purifying preferences of mistaken beliefs and non-self-regarding preferences.

A defender of the evidential defense may counter by saying that my argument is asking too much of the evidential defense—the platitudes that are used to purify preferences do not need to tell us *exactly* how to purify them, but merely need to provide a rough way that is good enough. If this is true, it is not clear that this rough way will do enough. Without a theory or a platitude telling me as a welfare economist how I should deal with default-biased preferences, what good is it to have a rough guide telling me that these preferences need to be purified? The level of specificity required for conducting any meaningful welfare analysis is more than is available by any rough guide or rule that we can call a generally-accepted platitude about well-being.

Finally, one might argue that there is a missing possibility in this argument—rather than purifying preferences or not purifying them, one could adopt the approach of Bernheim and Rangel (2009), which involves the analyst using only the consistent aspects of observed choices to reveal welfare-relevant preferences and treating welfare as indeterminate or unmeasurable in other cases. There is a debate over whether this methodology counts as preference purification. Bernheim (2021) argues that it does not, though it appears that Sugden disagrees (see Sugden, 2021: 422).[11] This distinction is not relevant for my argument here. Whether we call the approach preference purification, it does in fact involve departing from merely using all observed choices as inputs into a revealed preference welfare analysis.[12] My preceding arguments about the justification of the evidential defense and the practice or purifying preferences also apply to the Bernheim-Rangel methodology.

In summary, the existence of systematic behavioral anomalies creates problems for the evidential defense of welfare economics. If one attempts to use the evidential defense without purifying preferences of these behavioral anomalies, then one runs into serious problematic and implausible conclusions. If instead one attempts to use the evidential

defense after purifying preferences of behavioral anomalies, then this purification requires some theory of well-being more thorough than platitudes about what does and what doesn't count towards one's well-being.

Modifying the evidential defense to accommodate behavioral anomalies

I offer a brief sketch of how one could respond to these objections by revising the evidential defense in such a way as to allow it to accommodate behavioral anomalies. My claim here is that a possible way out is for the evidential defense to budge just a little bit on its assertion that it can get by without any theory of well-being. It need not have a full-fledged theory of well-being behind it, but it needs to have at least a notion of what types of preferences and behavior *cannot* constitute well-being. Here, I will develop this proposal in detail, providing a rough outline of a theory of welfare errors, which might guide future research on this issue, although it is beyond the scope of this paper to provide a full detailed defense of this proposal.

According to this argument, the evidential defense of welfare economics should be modified to assert that some common preferences and behaviors of people are *errors* and are clearly not preferences and behaviors that act to maximize their well-being. People can exhibit default bias, self-control problems, procrastination, distractions by irrelevant factors, and other features that render their preferences poor guides to their well-being. These types of preferences cannot be evidence for well-being, and they should not be used in welfare analysis. This modification adds another caveat or purification of preferences that are required of them before they can be used in welfare analysis: in addition to being self-interested and well-informed, preferences must be free of behavioral features that inhibit them of reflecting well-being. I call this the "modified evidential defense." My discussion in this section does not imply that the arguments criticizing the evidential defense presented in the previous section are made moot; the modified evidential defense is not a "solution" to the problems with the original evidential defense identified there. Rather, the discussion here offers a sketch of the argument for how revealed preference welfare economics could be justified even in the face of behavioral anomalies.

As I argued earlier, this is not possible to do without some theory of well-being. But, I now claim that it is not necessary to have a full-fledged theory of what *is* or what *constitutes* well-being; instead we require just some theory of some things that *are not* or *do not constitute* well-being. We can call this a theory of welfare *errors* rather than a theory of welfare.[13] For example, preferences that exhibit self-control issues, procrastination, or default bias are not good evidence of well-being because choices made with those preferences are generally not conducive to and do not reliably indicate improvements in well-being. It cannot be the case that what makes you better off systematically depends on what your default option was; it cannot be the case that what makes you better off systematically depends on the day of the week that you are making your plans. These statements appear to be uncontroversial, nevertheless to make these statements one needs some outline of a theory of errors that rules out some outcomes that cannot constitute well-being.

The claim that we require a theory of welfare errors may invite the question of how do we know, for example, that allowing a default to affect someone's choice is a welfare

error if we know nothing about what constitutes well-being. In fact, my argument rests on the assertion that it is not true that we know *nothing* about what constitutes well-being—it relies on the existence of a theory of welfare errors, though not a full-fledged theory of welfare. The original evidential defense makes the stronger claim that it can be used without knowing anything about what constitutes well-being, and as I have argued this claim is suspect.

This outline of a theory of welfare errors might even be called a set of platitudes, though we should be clear about what it is that we are talking about when we introduce platitudes. The concept of platitudes is vague in Hausman and McPherson, and I argue that it is not the case that the outline of a theory of welfare errors here is simply an extension of their reliance on platitudes. I take their notion of platitudes to mean statements that are uncontroversial to the point of banality or tautology, and that are unrelated to any theory of welfare. Their two claims fit this definition: that your preferences should be about yourself, and that your preferences should be well-informed, in order for them to provide evidence of well-being. As I argued in the previous section, the purification of behavioral anomalies from preferences does not fit this definition, so platitudes are insufficient. Their argument is clear in claiming to be agnostic about a theory of well-being; here what I am proposing is clear in claiming to *not* be agnostic about a theory of welfare errors. The purification of behavioral anomalies from preferences is distinct from and not merely an extension of the evidential defense's reliance on platitudes to purify preferences of mistaken beliefs and non-self-interested preferences. A theory of welfare errors does contain substantive, non-banal, non-tautological claims about welfare. These claims are required to salvage the evidential defense from the arguments presented in the previous section.[14]

A theory of welfare errors is easier to arrive at than is a theory of welfare. Here I do not need to provide a comprehensive list of errors, and I do not claim that the examples discussed earlier are exhaustive. I argue that it is less controversial to rule out specific cases where the satisfaction of preferences clearly does not promote well-being than it is to define what generally constitutes well-being. But one could still counter and argue that even the specific cases described here are not necessarily and unambiguously errors. Consider procrastination. One might suppose there is a chronic procrastinator who leads a happy and fulfilling life, and that his procrastination and impulsivity contribute to his happiness and fulfillment; what we call self-control problems may actually be constitutive of well-being for a dynamically-evolving person. Likewise, we must distinguish between preference changes that are not a welfare error (e.g. not going to the gym because you've broken your leg) from those that are welfare errors and should be purified (e.g. not going to the gym because you are procrastinating).

But the fact that these apparent errors may not be errors for some people does not negate the fact that they are errors for many other people, and for them struggles with procrastination and self-control negatively affect well-being. Procrastination need not be *universally* welfare-harming for it to be treated as generally a welfare error. Suppose that 90% of the instances of procrastination are errors that result in welfare reductions. Ten percent of procrastinators are the "happy procrastinators" described in the previous paragraph. If a welfare economist purified preferences of procrastination, she would be wrong 10% of the time. But if she didn't, she would be wrong 90% of

the time. It boils down to an empirical question about which type of procrastination is more common, and it seems likely that the welfare-error type is more common.

Regarding the distinction between legitimate or justifiable preference changes and ones that arise from procrastination, the theory of welfare errors would provide a justification for purifying the latter but not the former. Again, this paper does not attempt to provide a full detailed defense of such a theory, but I can speculate on how such a theory could go about providing the required distinction. A theory of welfare errors might claim that when preferences change over time in a way that is consistent with procrastination, but without a sufficient justification based on a change in relevant external circumstances, then those preferences are subject to being biased by welfare errors and can be purified. For instance, if it's the case that the timing at which a decision is made affects the decision in a way that is not welfare-relevant (as would be the case with present-biased preferences), than those preferences can be purified of this bias.

Similarly, I can provide a rough outline of how a theory of welfare errors would approach preferences that feature default bias. Such a theory might claim that preferences that depend on circumstances that are external and irrelevant to welfare like physical placement or defaults are subject to bias from welfare errors and can be purified. As an example, if the placement of food in a cafeteria (closer vs farther from arm's reach) affects an individual's choice, then the degree to which that affect choice can be justifiably purified from preferences on the basis of a theory of welfare errors. If the choice of a retirement plan is influenced by the default option offered to an employee, then that set of preferences can be purified from default bias, given a suitable theory of welfare errors that would justify it on grounds that such a default bias could not plausibly represent preferences that are reliably and consistently conducive to well-being. Again, the theory of welfare errors would have to delineate and defend these claims about welfare errors, which is something that I am not doing here, and doing so would not be agnostic about a theory of welfare as the original evidential defense claims to be.

Given this outline of a theory of well-being errors, how does the modified evidential defense operate? It need not go so far as to dictate the form of the purified utility function, or to dictate exactly how preferences are to be purified. But, it can still defend the *practice* of using a particular purified utility function on the grounds that the welfare evaluated with that purified utility function is good evidence for well-being. While the original version of the evidential defense claims that preferences provide the best available evidence of well-being although they do not constitute well-being, the modified evidential defense might claim that a purified utility function is good evidence of well-being though it does not define or constitute well-being.

This comparison brings up one additional distinction between the original evidential defense and the modified evidential defense. The original evidential defense explicitly claims that preferences are the *best available* evidence of welfare, not merely good evidence.[15] The modified evidential defense will not necessarily be able to claim that (sufficiently purified) preferences are the best available evidence of welfare, but merely good evidence. With just a theory of welfare errors, it will be difficult or impossible to arrive at a justification that a particular measure of welfare is better than any potential other measure. This does not mean that the modified evidential defense is

worthless, since it still provides a justification for using preferences to measure welfare, but it cannot rule out the existence of better evidence.

In arguing for such a defense of behavioral welfare economics, we need not be wedded to any particular purified utility function or any preference purification strategy. The modified evidential defense does not need to identify the sole strategy that ought to be used in behavioral welfare economics. It can defend several ways to purify preferences of their behavioral anomalies. An argument analogous to that of Hersch (2015) might object to this lack of specificity of the evidential defense; Hersch (2015) argued that other evidence besides preferences and choices could be used in measuring well-being, like subjective well-being surveys. But here the lack of specificity is not detrimental to the modified evidential defense—this modified evidential defense is still making the claim that preferences give us good evidence of what promotes well-being, it just does not specify how exactly preferences are to be purified, so long as the purification method rids preferences of the features that are unambiguously unrelated to things that can promote our well-being.

Acknowledgements

The author thanks Spencer Banzhaf, Andrew J. Cohen, Glenn Harrison, Gil Hersch, Yongsheng Xu, and seminar participants at GSU and at the PPE Society conference for helpful comments.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Garth Heutel https://orcid.org/0000-0002-4059-2675

Notes

1. The argument is also presented in Hausman (2012: 88–93).
2. Throughout this paper I will use "revealed preference" and "revealed preference welfare economics" interchangeably to refer to this interpretation that choices indicate preferences and preferences indicate welfare. This is a standard interpretation of the term: "A common precept of standard economics is that people only make the choices that maximize their welfare. This assumption even has a fancy name, 'revealed preferences': that people reveal what makes them better off by their choices" (Akerlof and Shiller, 2015: 170). This is distinct from Samuelson's "revealed preference theory," which is only about the link between observed

choices and inferring preferences and has nothing to do with whether welfare is observed from preferences. For earlier discussion of revealed preference theory and its possible normative implications, see Sen (1973) and Anderson (2001).

3. The economic and philosophical literature on behavioral welfare economics is broad, and I will not attempt here to summarize it in its entirety. Several recent papers on the philosophy of economics address issues related to behavioral welfare economics (BWE). Mitrouchev and Buonomo (forthcoming) study the relationship between BWE and identity/multiple selves. Abrahamson (forthcoming) explores the importance of the source of context-dependence in preferences. For more thorough discussions, see Bernheim and Taubinsky (2018) or Rizzo and Whitman (2020).
4. See DesRoches (2020) for a discussion of the justification of the distinction between these two types of utility functions and two types of preferences.
5. Later I will address the issue of whether the methodology proposed by Bernheim and Rangel (2009) offers a third option, which is neither purifying preferences nor taking them as given.
6. Whitman and Rizzo (2015) address these questions as well but find little justification for any particular method of preference purification.
7. The abstract in Hausman and McPherson (2009) claims their evidential defense "is independent of any philosophical theory of well-being" (p. 1).
8. Researchers have verified present bias among some subjects in many laboratory experiments and real-world situations, for example, DellaVigna and Malmendier (2006), Benhabib et al. (2010), and Montiel Olea and Strzalecki (2014).
9. Ancient Greek philosophers wrote extensively about weakness of will or akrasia (Bobonich and Destrée 2007).
10. This critique of the evidential defense is reminiscent of some discussion in Hersch (2015). He argues that "relying on platitudes fails to uniquely justify relying on choices as evidence for what is conducive to well-being" (p. 285). Hersch lists other sources of evidence for well-being, like subjective well-being surveys, that might provide better evidence than choices and are arguably just as justified as choices are. Here, I am arguing that there seems to be no justification for purifying preferences of their behavioral anomalies without a theory of well-being and that the appeal to platitudes cannot be invoked.
11. The argument cited specifically concerns whether the approach relies on the notion of an "inner rational agent," criticized in Infante et al. (2016).
12. Bernheim (2021: 391) acknowledges that the method "requires us to identify the circumstances under which the individual suffers from characterization failure."
13. My proposal is similar in spirit to that of Hersch (2020), who claims that a completely theory-free account of well-being that can be used in policy guidance is impossible. Instead, he proposes an "intermediate account," which refers to substantive well-being theories but in a way as agnostically as possible. See also Grüne-Yanoff (2022), who argues for an "algebraic" interpretation of preferences to be consistent with behavioral welfare economics. Finally, this is also reminiscent of the basis of the methodology in Bernheim (2021); see his "Step 1" on p. 90 and his justification for identifying errors.
14. One can compare this to the discussion in Hausman (2020: 13–16), who offers what he calls a "folk theory of well-being," which relies on "some weak premises concerning what conduces to well-being to identify factors that sometimes make preferences poor indicators of well-being."

15. For instance, the authors claim "Regardless of what philosophical theory of human well-being one accepts, the best indicator of well-being in certain circumstances is the extent to which preferences are satisfied." (Hausman and McPherson, 2009: 18). And, "Even though what satisfies Ann's preferences does not necessarily make her better off, Ann may be sufficiently self-interested and well informed that her preferences are the best guide others have to what is beneficial to her. What better way is there to determine what will benefit people?" (p. 16).

References

Abrahamson M (2024) Permissible preference purification: on context-dependent choices and decisive welfare judgements in behavioural welfare economics. *Journal of Economic Methodology* 31(1): 17–35.

Akerlof GA and Shiller RJ (2015) *Phishing for Phools: The Economics of Manipulation and Deception*. Princeton: Princeton University Press.

Anand P (1987) Are the preference axioms really rational? *Theory and Decision* 23(2): 189.

Anderson E (2001) Symposium on Amartya Sen's philosophy: 2 unstrapping the straitjacket of 'preference': a comment on Amartya Sen's contributions to philosophy and economics. *Economics & Philosophy* 17(1): 21–38.

Angner E and Loewenstein GF (2012) Behavioral economics. In: Mäki U (ed) *Handbook Of The Philosophy Of Science: Philosophy Of Economic*. Amsterdam: Elsevier, pp. 641–690.

Benhabib J, Bisin A, and Schotter A (2010) Present-bias, quasi-hyperbolic discounting, and fixed costs. *Games and Economic Behavior* 69(2): 205–223.

Bernheim BD (2021) In defense of behavioral welfare economics. *Journal of Economic Methodology* 28(4): 385–400.

Bernheim BD and Rangel A (2009) Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *The Quarterly Journal of Economics* 124(1): 51–104.

Bernheim BD and Taubinsky D (2018) Behavioral public economics. In: *Handbook of Behavioral Economics: Applications and Foundations 1*, vol. 1. North-Holland: North-Holland, pp. 381–516.

Bleichrodt H, Pinto JL, and Wakker PP (2001) Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science* 47(11): 1498–1514.

Bobonich C and Destrée P (eds) (2007) *Akrasia in Greek Philosophy: From Socrates to Plotinus*, vol. 106. Leiden: Brill.

Choi JJ, Laibson D, Madrian BC, et al. (2004) For better or for worse: default effects and 401 (k) savings behavior. In: *Perspectives on the Economics of Aging*. Chicago: University of Chicago Press, pp. 81–126.

DellaVigna S and Malmendier U (2006) Paying not to go to the gym. *American Economic Review* 96(3): 694–719.

DesRoches CT (2020) Value commitment, resolute choice, and the normative foundations of behavioural welfare economics. *Journal of Applied Philosophy* 37(4): 562–577.

Dold MF and Rizzo MJ (2021) The limits of opportunity-only: context-dependence and agency in behavioral welfare economics. *Journal of Economic Methodology* 28(4): 364–373.

Grüne-Yanoff T (2022) What preferences for behavioral welfare economics? *Journal of Economic Methodology* 29(2): 153–165.

Hausman DM (2012) *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.

Hausman DM (2020) Enhancing welfare without a theory of welfare. *Behavioural Public Policy* 6(3): 342–357.

Hausman DM and McPherson MS (2009) Preference satisfaction and welfare economics. *Economics & Philosophy* 25(1): 1–25.

Hersch G (2015) Can an evidential account justify relying on preferences for well-being policy?. *Journal of Economic Methodology* 22(3): 280–291.

Hersch G (2020) No theory-free lunches in well-being policy. *The Philosophical Quarterly* 70(278): 43–64.

Infante G, Lecouteux G, and Sugden R (2016) Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology* 23(1): 1–25.

Johnson EJ, Hassin R, Baker T, et al. (2013) Can consumers make affordable care affordable? The value of choice architecture. *PloS One* 8(12): e81521.

Lahiri S (2019) Default bias in extended choice rules. *Studies in Microeconomics* 7(1): 1–6.

Mahmoud O (2017) On the consistency of choice. *Theory and Decision* 83(4): 547–572.

Mitrouchev I and Buonomo V (forthcoming) Identity, ethics and behavioural welfare economics. *Economics and Philosophy*: 1–27.

Montiel Olea JL and Strzalecki T (2014) Axiomatization and measurement of quasi-hyperbolic discounting. *The Quarterly Journal of Economics* 129(3): 1449–1499.

Mullainathan S and Thaler RH (October 2000) Behavioral economics. NBER Working Paper No. w7948.

O'Donoghue T and Rabin M (2006) Optimal sin taxes. *Journal of Public Economics* 90(10–11): 1825–1849.

Rizzo MJ and Whitman G (2020) *Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy*. Cambridge: Cambridge University Press.

Sarch AF (2015) Hausman and McPherson on welfare economics and preference satisfaction theories of welfare: a critical note. *Economics & Philosophy* 31(1): 141–159.

Sen A (1973) Behaviour and the concept of preference. *Economica* 40(159): 241–259.

Sugden R (2018) *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford: Oxford University Press.

Sugden R (2021) A response to six comments on *The Community of Advantage*. *Journal of Economic Methodology* 28(4): 419–430.

Thoma J (2021) On the possibility of an anti-paternalist behavioural welfare economics. *Journal of Economic Methodology* 28(4): 350–363.

Whitman DG and Rizzo MJ (2015) The problematic welfare standards of behavioral paternalism. *Review of Philosophy and Psychology* 6(3): 409–425.