Behavioral Economics and the Evidential Defense of Welfare Economics

Garth Heutel[1]

Georgia State University

gheutel@gsu.edu

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.

[1] Thanks to 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.

I. 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 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.[2] 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 differs 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 biases or anomalies, and these anomalies create a systematic wedge between their

[2] The argument is also presented in Hausman (2012, p. 88-93).

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 III) 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 present 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 section IV, 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.

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. Section III presents my main argument, which is that the evidential defense encounters problems in dealing with behavioral anomalies. I conclude (section IV) with a brief sketch of a proposal for how the evidential defense could be modified to attempt to address these concerns.

II. 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."[3] 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

[3] 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, p. 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).

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."

> If people are more or less self-interested with respect to certain alternatives, then economists can use people's preferences to make inferences concerning what people believe will benefit them. And if it is also reasonable to suppose with respect to the policies being considered and their consequences that individuals are good judges of what will benefit them, then economists can use people's preferences as evidence concerning what in fact makes them better off. In this way welfare economists can defend their practice of making inferences concerning well-being from people's preferences without committing themselves to any theory of well-being at all. (2009: 16)

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. It merely relies on the assertion that individuals generally know what will improve their welfare and generally make

choices and have preferences to maximize their welfare, so that performing revealed preference welfare analysis will generally be a good guide to measuring well-being.

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. The issue of what to do with non-self-interested preferences still arises. These types of preferences do not seem like they truly affect well-being, but the evidential defense of welfare economics suggests that they ought anyway to be seen as evidence for well-being. Hausman and McPherson (2009) propose that the welfare economist should not concern herself with preferences that are not related to her own judgment about what is better for her.

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. While this appeal to platitudes (which could also be interpreted as intuition) is somewhat unsatisfying and seemingly incomplete, this evidential view is inherently agnostic about what constitutes well-being and is not an attempt to provide such a theory.

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. It is a collection of observed "behavioral anomalies" or "behavioral biases" that demonstrate how people deviate from rational choice theory. Researchers have identified several such anomalies and have

developed several descriptive theories of behavior to accommodate those anomalies. In this paper, I will use one specific anomaly – present bias – as an example to help illuminate my argument.

The standard neoclassical model asserts that people make intertemporal decisions using a constant discount rate. A significant body of evidence shows that some people do not make intertemporal decisions with a constant, time-consistent discount rate. Instead, people exhibit present bias – placing extra value on the present period relative to all other periods. This bias implies that intertemporal decision-making cannot be modeled with a single discount rate. Today, when I evaluate consumption ten years into the future relative to consumption eleven years into the future, I use a 5% discount rate. But today, when I evaluate today's consumption relative to next year's consumption, I use a 10% discount rate. Economists model this type of present bias using quasi-hyperbolic preferences (Laibson 1997), where the discount factor used between any two consecutive future periods is δ, while the discount factor used between now and the following period is $\beta \times \delta$, where $\beta < 1$ (quasi-hyperbolic discounting is also called "$\beta\delta$ discounting").[4]

What are the implications of behavioral economics for welfare economics? That topic is a broad area that I can only briefly summarize.[5] If people's behavior cannot be modeled by rational choice theory, then the link between observed choice and well-being, at the heart of the justification for revealed preference welfare analysis, must be called into question. Much of the literature devoted to behavioral welfare analysis has taken the approach that has been called "preference purification" (Hausman 2012, Infante et al. 2016). Preferences determine how

[4] 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 et al. (2014).
[5] For more thorough discussions, see Bernheim and Taubinsky (2018) or Rizzo and Whitman (2020).

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 ways in which 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. I will refer to the two utility functions as *decision utility* and *purified utility*.

Present bias offers a good opportunity for describing how policy analysis and policy design are conducted in the decision utility vs. purified utility framework. O'Donoghue and Rabin (2006) conduct welfare analysis in a model of consumers with present bias modeled by quasi-hyperbolic ($\beta\delta$) preferences. The decision utility function includes both the present bias discount factor β and the long-run discount factor δ. However, the purified utility function includes only the long-run discount factor δ. The interpretation is that the present bias β represents self-control problems, procrastination, or some other psychological factor that prevents people from acting in a way that maximizes their true well-being. The decision utility function is given by the quasi-hyperbolic specification. However, the purified utility function, or the individual welfare function, is given by the constant discount rate specification. That is, the purified utility function imposes that the present bias discount factor β be set equal to one.[6]

[6] Economists have also used this method in the context of environmental policy in Heutel (2015), in the context of Kenyan farmers investing in fertilizer in Duflo et al. (2011), and in the context of enrollments in 401(k) retirement accounts in Carroll et al. (2009).

III. The Evidential Defense in Light of Behavioral Economics

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 a mistake 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.

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 present-biased preferences, where someone uses a higher discount rate between today and tomorrow than he does between tomorrow and the day after tomorrow. There is nothing mistaken in this behavior – it isn't that the person is wrong about what day it is. Rather, the preferences are well-informed but generate the behavioral anomaly of time inconsistency. 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 biases, so long as those preferences are well-informed and self-interested. By way of example consider present-biased preferences again. On Sunday, my preferences over when I will go to the gym this week are that I go on Tuesday and Thursday. But on Tuesday, I prefer instead to go on Wednesday and Thursday. Then, on Wednesday, I prefer to go on Thursday and Friday.[7] Each of these preferences is self-interested – they are about me going to the gym for my own benefit. And each of these preferences is well-informed – nowhere am I mistaken about what day it is or about the effects of going or not going to the gym. If using the evidential defense without purifying preferences of their time-inconsistency, one is claiming that the best evidence for what maximizes my well-being on Sunday is going to the gym on Tuesday and Thursday, but the best evidence for it on Tuesday is going to the gym on Wednesday and Friday. This seems very unsatisfactory. Preferences may be time-inconsistent in this way, but it does not seem plausible that *what constitutes well-being* is similarly time-inconsistent. It cannot be that, evaluated on Sunday, what promotes my best interests or well-being for the week is a

[7] This pattern of preferences can be modeled as a series of intertemporal decisions by an agent with quasi-hyperbolic ($\beta\delta$) preferences, where going to the gym yields costs (negative utility) now but benefits in the future (DellaVigna and Malmendier 2006).

substantially different course of action than what promotes my best interests or well-being for the week, evaluated on Tuesday, when the only difference is the day of the week that I am doing the evaluating.

This argument does not rule out the possibility that preferences can rationally change. If something else is different between Sunday and Tuesday that is plausibly welfare-relevant, then things are different. For example, if I break my leg on Monday and am unable to attend the gym afterward, then it is reasonable to assert that the course of action that best promotes my well-being has indeed changed; likewise if the gym unexpectedly raises its admissions price between Sunday and Tuesday. What makes a preference change exemplify present bias or time inconsistency is when nothing welfare-relevant changes that justifies the preference reversal.

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. Ask a welfare economist for a cost-benefit analysis on a Monday, and you'll get a different evaluation than if you ask her on a Wednesday. 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?

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 present 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.[8] 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

[8] The abstract in Hausman and McPherson (2009) claims their evidential defense "is independent of any philosophical theory of well-being" (p. 1).

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, i.e. 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. Consider time inconsistency, which leads to present-biased preferences. This behavioral anomaly can explain (under one interpretation) the act of procrastination. As I described earlier, decisions made under

procrastination do not seem to be reliable indicators of what promotes well-being, and 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. For example, someone prefers yesterday to go to the gym today, but today prefers to wait until tomorrow (where nothing else has changed that would justify a rational preference change). A part of their preference for not going today contributes to their welfare, since it avoids the pain or exertion that comes with exercising. But (re-phrasing Hausman and McPherson's quote), it is hard to see what other contribution to individual welfare the delaying of going to the gym could make, because it is hard to see how the delay bears on other intrinsic goods, like health. 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. Under this argument, the point that I am making here about purifying preferences of their behavioral anomalies is not a critique of the evidential defense but simply a good addition to their main argument. That is, the burden is not on the evidential defense to have to defend this third form of purification, any more than it has to defend the uncontroversial first two forms of purification.

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, e.g. 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. Consider a different example of a behavioral anomaly: default bias. 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. 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 *mistaken* – e.g., when you prefer to skip going to the gym you are making a mistake about what is in your own best interest. One could possibly argue that this claim about preferences being mistaken (not mistaken about objective facts of the world, but mistaken 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 mistakes 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

[9] Ancient Greek philosophers wrote extensively about weakness of will or akrasia (Bobonich and Destrée 2007).

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 present bias caused by time-inconsistent preferences. The standard way of purifying preferences in this case is to assert that the present bias discount factor β is welfare-irrelevant and to only use the long-run discount factor δ in welfare analysis (O'Donoghue and Rabin 2006). 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 (in particular, a theory of well-being that states that only the time-consistent part of preferences is related to well-being, or that the welfare function must exhibit time-consistent preferences). 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.

It is perhaps even harder to justify a specific way to purify preferences in the example of default bias. The closest thing here 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. 2011). I can think of no platitudes

[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.

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).

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; i.e. 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 errors. These two groups of things are not identical to each other, though they may appear to be. Uncontroversial welfare errors include those emerging from time-inconsistent preferences. It cannot be conducive to my well-being on Sunday to go to the gym on Monday, but then on Monday no longer conducive to my well-being to go on Monday. 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 time-inconsistency 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. This is the heart of my argument in this paper.

Allow me to expand on this distinction. Consider again the specific example where on Sunday I plan to go to the gym on Monday, but when Monday arrives I plan instead to delay until Tuesday. This is unreliable evidence for well-being, since the day of the week in 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 time-inconsistency, 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 time-inconsistent 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.

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.

IV. 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.

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 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.[11] 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

[11] 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.

makes you better off systematically depends on the day of the week that you are making your plans; it cannot be the case that what makes you better off systematically depends on what your default option was. 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.[12]

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 (for example, not going to the gym because you've broken your leg) from those that are welfare errors and should be purified (for example, 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-

[12] One can compare this to the discussion in Hausman (2020, p. 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."

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. 10% 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. While in this paper I do not attempt to provide and defend such a theory, 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. Likewise in the case of preferences that are influenced by framing effects, a theory of welfare errors 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. 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.[13] 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

[13] 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, p. 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).

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.

References

Akerlof, G. A., & Shiller, R. J. (2015). *Phishing for phools: The economics of manipulation and deception*. Princeton University Press.

Anderson, Elizabeth. "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, no. 1 (2001): 21-38.

Benhabib, Jess, Alberto Bisin, and Andrew Schotter. "Present-bias, quasi-hyperbolic discounting, and fixed costs." *Games and Economic Behavior* 69, no. 2 (2010): 205-223.

Bernheim, B. Douglas, and Dmitry Taubinsky. "Behavioral public economics." In *Handbook of Behavioral Economics: Applications and Foundations 1*, vol. 1, pp. 381-516. North-Holland, 2018.

Bobonich, Christopher, and Pierre Destrée, eds. *Akrasia in Greek Philosophy: From Socrates to Plotinus*. Vol. 106. Brill, 2007.

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick. "Optimal defaults and active decisions." *The Quarterly Journal of Economics* 124, no. 4 (2009): 1639-1674.

Choi, James J., David Laibson, Brigitte C. Madrian, and Andrew Metrick. "For better or for worse: Default effects and 401 (k) savings behavior." In *Perspectives on the Economics of Aging*, pp. 81-126. University of Chicago Press, 2004.

DellaVigna, Stefano, and Ulrike Malmendier. "Paying not to go to the gym." *American Economic Review* 96, no. 3 (2006): 694-719.

Duflo, Esther, Michael Kremer, and Jonathan Robinson. "Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya." *American Economic Review* 101, no. 6 (2011): 2350-90.

Hausman, Daniel M. *Preference, value, choice, and welfare*. Cambridge University Press, 2012.

Hausman, Daniel M. "Enhancing welfare without a theory of welfare." *Behavioural Public Policy* (2020): 1-16.

Hausman, Daniel M., and Michael S. McPherson. "Preference satisfaction and welfare economics." *Economics & Philosophy* 25, no. 1 (2009): 1-25.

Hersch, Gil. "Can an evidential account justify relying on preferences for well-being policy?." *Journal of Economic Methodology* 22, no. 3 (2015): 280-291.

Hersch, Gil. "No theory-free lunches in well-being policy." *The Philosophical Quarterly* 70, no. 278 (2020): 43-64.

Heutel, Garth. "Optimal policy instruments for externality-producing durable goods under present bias." *Journal of Environmental Economics and Management* 72 (2015): 54-70.

Infante, Gerardo, Guilhem Lecouteux, and Robert Sugden. "Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics." *Journal of Economic Methodology* 23, no. 1 (2016): 1-25.

Johnson, Eric J., Ran Hassin, Tom Baker, Allison T. Bajger, and Galen Treuer. "Can consumers make affordable care affordable? The value of choice architecture." *PloS one* 8, no. 12 (2013): e81521.

Laibson, David. "Golden eggs and hyperbolic discounting." *The Quarterly Journal of Economics* 112, no. 2 (1997): 443-478.

Montiel Olea, José Luis, and Tomasz Strzalecki. "Axiomatization and measurement of quasi-hyperbolic discounting." *The Quarterly Journal of Economics* 129, no. 3 (2014): 1449-1499.

O'Donoghue, Ted, and Matthew Rabin. "Optimal sin taxes." *Journal of Public Economics* 90, no. 10-11 (2006): 1825-1849.

Pinto-Prades, Jose-Luis, and Jose-Maria Abellan-Perpiñan. "When normative and descriptive diverge: how to bridge the difference." *Social Choice and Welfare* 38, no. 4 (2012): 569-584.

Rizzo, Mario J., and Glen Whitman. *Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy*. Cambridge University Press, 2020.

Sarch, Alexander F. "Hausman and McPherson on welfare economics and preference satisfaction theories of welfare: a critical note." *Economics & Philosophy* 31, no. 1 (2015): 141-159.

Sen, Amartya. "Behaviour and the Concept of Preference." *Economica* 40, no. 159 (1973): 241-259.