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 revealed preference 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 propose a slight modification of the evidential defense that allows it 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 revealed preference: 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, including cost-benefit analyses, is on shaky ground.

Hausman and McPherson (2009) present an *evidential defense* of the standard methodology of conducting welfare economics by using revealed preferences.[2] According to their argument, even though preferences are not *constitutive* of welfare, they nonetheless provide *evidence* for what things promote welfare, and perhaps they provide the best available evidence, given heterogeneity across people in what promotes welfare and regulators' imperfect information. Thus, the standard method of conducting welfare analysis in neoclassical economics, based on revealed preference, is justified on grounds of pragmatism.

Economists and psychologists have, over the past several decades, developed the field of behavioral economics, which, loosely speaking, points out that the way that some people act is often not the way that they are modeled to act in standard economic models. These people's deviations from the standard models' predictions, or their "behavioral anomalies," are not random but are biased in certain directions. The existence of these behavioral anomalies also presents problems for the standard revealed preference methodology of welfare economics: if the

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

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 must be purified of their behavioral anomalies before they can be used to assess welfare. This methodology is based on the assertion that behavioral anomalies introduce a distinction between the actions that people take and the actions that would in fact increase their welfare. In some instances, this has been termed a distinction between one's *decision utility* and one's *experienced utility*. The decision utility governs how people behave and make decisions, while the experienced utility (also sometimes called *hedonic utility*) describes the things that truly increase people's welfare and make them better off. A benevolent planner (or an economist conducting welfare analysis) ought to be concerned with experienced utility, and, therefore, should purify preferences to transform the decision utility function into the experienced utility function.

The evidential defense of welfare economics runs into problems when revealed choices themselves do not give very good evidence of welfare. When individuals are subject to behavioral biases or anomalies, and these anomalies create a systematic wedge between decision and experienced utility, then claiming that revealed preference can guide welfare analysis inappropriately confounds decision utility with experienced utility. The evidential defense runs into difficulties when a preference purification method is applied to attempt to rid preferences of their behavioral anomalies and allow them to reflect well-being.

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 IV) 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 constitutes well-being, which the evidential defense purports to be agnostic about.

I therefore propose a solution (presented in detail in section V), which allows 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. While it does not require a full-fledged theory of well-being, it at least requires a theory describing what kinds of actions and preferences *do not* provide consistent evidence for 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, I briefly introduce several problems with it that have been identified in the literature, and I summarize the evidential defense of welfare economics, which seeks to overcome many of these problems. I also discuss some previous papers that have objected to the evidential defense on other grounds. In section III, I briefly review the main findings of the field of behavioral economics and describe two specific behavioral anomalies (present bias and loss aversion). I also discuss behavioral economics' implications for welfare economics and the preference purification approach. Readers who are familiar with the standard practice of welfare economics, its objections, and the evidential defense may wish to skip section II; readers who are familiar with behavioral welfare

economics and preference purification may wish to skip section III. Section IV presents my main argument, which is that the evidential defense encounters problems in dealing with behavioral anomalies. I conclude (section V) with a proposal for how the evidential defense can be modified to attempt to address these concerns.

II. Welfare Economics and the Evidential Defense

In this section I briefly discuss the standard interpretation of welfare economics, which is commonly used by welfare economists conducting cost-benefit and other policy analyses. It is based on the notion of revealed preference, which is that the choices that people make allow observers to infer their preferences and that satisfying their preferences is constitutive of maximizing their welfare or well-being. I will then discuss some common objections, which in part motivate the evidential defense. I end this section summarizing the evidential defense.[3]

Economists are often concerned about the effects of policy or of markets on people's welfare or well-being.[4] How can economists (or anyone) know these effects on people's welfare? 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" – by observing peoples' actual choices, their preferences are being revealed to the economist.[5] 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. Therefore, when conducting welfare analysis, a policy

[3] For earlier discussion of revealed preference theory and its normative implications, see Sen (1971, 1973) and Anderson (2001) for a comment on Sen's philosophy in this area.
[4] Throughout this paper I will use the terms "welfare" and "well-being" interchangeably.
[5] Hausman (2011) discusses what is meant by "preferences." Roughly, preferences are an ordering or a ranking of states of the world; we can say that a person i prefers state X to state Y, meaning that if both are available, then i would choose to have state X rather than state Y.

that can induce X will yield higher welfare than a policy that can induce Y, all else equal (i.e. assuming there are no differences across the policies or the states that affect anyone else's welfare). One interpretation of this method is that welfare economists are in fact equating preference satisfaction with well-being: for something to increase or promote my well-being is equivalent to it better satisfying my preferences. If I prefer X to Y, then it must be the case that X gives me higher welfare than does Y. This is the preference-satisfaction theory of well-being.

The preference-satisfaction theory of well-being is implicitly or explicitly utilized in most actual welfare economics studies. Consider a cost-benefit analysis (CBA), which tallies up all of the costs associated with a proposed or actual policy and all of the benefits, then typically presents values for both the total costs and the total benefits. Implicitly (or explicitly) it claims that a policy is good, or increases social welfare, if the total benefits exceed the total costs. Notably, in almost all cost-benefit analyses, the costs and benefits are calculated based on a revealed preference methodology, so that they are equated with preferences as described above.[6]

There are several well-known problems with the preference-satisfaction theory of well-being, and therefore with using it to justify revealed preference welfare economics like cost-benefit analyses. It seems rather obvious that people sometimes make "bad choices," i.e. they prefer something that does not increase their welfare. People can be addicted to drugs, food, or video games, or they can be chronic procrastinators. In each case their preferences do not align

[6] We can interpret this in relation to the first fundamental theorem of welfare economics, which states that market equilibrium is Pareto efficient under certain conditions. The purpose of a CBA can be seen as justifying a policy intervention correcting for an inefficient market allocation when at least one of the conditions behind the welfare theorem is not met, but in a way that satisfies a market-like test.

with their welfare; satisfying their preferences will not increase their welfare. This phenomenon is incompatible with the preference-satisfaction theory of welfare.[7]

Bad choices can arise for several reasons (Reiss 2013, p. 216). First, people can be misinformed. I might believe that spending hundreds of dollars on a weight-loss scheme will help me lose weight and therefore prefer to spend the money on the scheme rather than spend it elsewhere or save it. However, it might be the case that the scheme is useless, and so undertaking the scheme will not cause me to lose weight and will not increase my welfare. Second, people can fail to act on their well-informed preferences, for instance through weakness of will or self-control problems. An entire industry exists to combat such psychological barriers to acting in accordance with our will, some of which incidentally use findings from behavioral economics to overcome these barriers.[8] Third, people may have preferences that do not relate to their well-being, including altruistic preferences. This third set of violations of the preference-satisfaction theory of well-being does not contain "bad" choices in a moral sense (nor necessarily do the other two sets), but bad insofar as one cannot equate preference satisfaction with increasing well-being.

Because of the aforementioned philosophical and practical concerns with the preference-satisfaction theory of well-being and with revealed preference welfare economics, 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-

[7] Harrison and Ng (2016) provide laboratory evidence for "bad choices" – people make decisions over insurance that fail to maximize their well-being.
[8] For example, see the website stickk.com.

benefit analyses, and treating preferences as if they were constitutive of welfare is a defensible, practical strategy for conducting welfare analysis. I now summarize their argument, which I will call the "evidential defense of welfare economics," or just the "evidential defense."

Before introducing their evidential defense, Hausman and McPherson (2009) discuss another potential remedy for the issues with the preference-satisfaction theory of welfare, which is the notion of "laundered" preferences. Not all preferences can be equated with well-being, but only certain types of preferences or preferences that satisfy certain conditions. These "properly pruned and purified preferences" (Hausman and McPherson 2009, p. 11) are generally thought to have to satisfy two conditions: they must be self-interested and informed.

The laundered preferences approach is rejected by Hausman and McPherson (2009), because it still relies on the satisfaction of a set of preferences, albeit laundered preferences, being definitive of well-being. "The problem… is that the fact that Ben prefers *x* to *y* does not make it the case that *x* is better for him than *y*, no matter what condition one imposes on his preferences… Preference satisfaction has by itself nothing to do with well-being" (p. 12). It is impossible, they argue, to claim that sufficiently laundered preferences are definitive of well-being without relying on some theory of well-being that is not based on preference satisfaction. For example, if the laundered preferences increase well-being because one is pleased at their satisfaction, whereas non-laundered preferences being satisfied does not please, then it is the *pleasure* from the preference satisfaction that makes it welfare-increasing rather than the preference satisfaction itself.

With this in mind, Hausman and McPherson (2009) move on to propose their defense of welfare economics based on the evidence for well-being provided by preference satisfaction.

> 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. Preferences are sometimes good evidence for welfare. (2009: 16)

The power of this defense lies in its agnosticism about a theory of well-being (i.e. a theory of what constitutes well-being). It emphatically does not equate well-being with preference satisfaction, nor does it imply that well-being is somehow close to preference satisfaction or related to preference satisfaction. Instead, it merely relies on the assertion that individuals generally know what it is that 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 actual well-being.

Furthermore, revealed preference welfare economics is likely the *best* practical way to conduct welfare economics. Even if the link between preferences and well-being is weaker than we would like, so that the evidential link between the two is fuzzy, preferences may still be the best and strongest evidence that economists (or anyone) have about people's well-being. Since people are different from each other, different preferences are indicative of different determinants of well-being. Objective criteria for determining well-being (like Nussbaum's

(2003) ten "central human capabilities") fail to appreciate the diversity across people. A revealed preference approach is the best way to accommodate this diversity, according to their argument.

It is clear (and acknowledged by Hausman and McPherson (2009)) that the evidential defense does not do away with all of the standard objections to revealed preference and the preference-satisfaction theory of well-being. In particular, the issue of what to do with non-self-interested preferences still arises. These types of preferences, which include for example "existence values" for endangered species, 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. The escape from this conundrum offered by Hausman and McPherson (2009) is somewhat analogous to the laundered preferences approach discussed earlier – the welfare economist should not concern herself with preferences that are not related to her own judgment about what is better for her. This is not quite the same as ignoring non-self-interested preferences, since such preferences may nevertheless be about one's judgement of one's well-being. For example, a preference to preserve an endangered species is non-self-interested and probably not about one's well-being, but a preference that one's child not be convicted of a crime is non-self-interested but probably related to one's well-being.[9]

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

[9] See also Bergstrom (2003) on the issue of double-counting well-being among people with altruistic preferences.

judgment as to what preferences provide evidence for well-being; after all, their entire justification for welfare economics rests on the supposition that individuals are using their best judgment when forming their preferences. While this appeal to platitudes (which could also be interpreted as intuition, or one's best judgment) is somewhat unsatisfying and seemingly incomplete, it is important to bear in mind that this evidential view is inherently agnostic about what constitutes well-being and is not an attempt to provide such a theory. Instead, it is a practical defense of the use of revealed preference welfare economics. The fact that it appeals to platitudes in its attempt to launder allowable preferences does not bring the entire theory crashing down, since it is not a theory (of well-being) in the first place.

Hersch (2015) provides a critique of the evidential defense. Hersch (2015) claims that the evidential defense does not uniquely justify the revealed preference approach to welfare economics without relying (as it claims not to do) on some specific theory of well-being. The evidential defense is based on three assumptions, according to Hersch. First, people must have true beliefs about feasible alternatives; second, people must be competent evaluators about what affects their well-being; and third, choices must be aimed at maximizing one's own well-being. Without a theory of well-being, we have no way of knowing whether or not the second or third assumptions hold. The response of Hausman and McPherson is the appeal to platitudes – that platitudes give us reasons for why individuals' preferences are related to their well-being. Hersch's (2015) objection is to point out that platitudes do not uniquely justify revealed preference: "platitudes can inform us of when the assumptions that justify viewing *several* measures as evidence for what is conducive to well-being hold… relying on platitudes fails to uniquely justify relying on choices as evidence for what is conducive to well-being." (p. 285, emphasis in original). For example, other measures of well-being, like subjective well-being

(SWB) surveys or objective measures, like those in the Human Development Index, can also be interpreted as having platitudes justify their use for revealing information about well-being. A welfare economist may prefer to conduct a welfare analysis based on SWB surveys[10] instead of based on revealed preference, and there is no reason offered by Hausman and McPherson (according to Hersch) for why revealed preference should be the preferred strategy. The fact that individuals are generally competent evaluators of their own welfare (one of the conditions required for the evidential defense of welfare economics) also seems to imply that a SWB-based welfare analysis is also justified, perhaps more so than a revealed preference analysis.[11]

Finally, it may seem that this entire set of arguments takes place against an assumption of a welfarist moral theory. Whether one supports the practice of revealed preference or the evidential defense of it, the arguments presented here and in Hausman and McPherson (2009) are about welfare and how to measure welfare to use in policy guidance. For somebody who rejects a welfarist moral theory, does this debate have any relevance? I think that the answer is yes, because even those who do not subscribe to a welfarist moral theory may (and perhaps should) nonetheless care about human well-being.[12] For example, a deontological ethicist may still care about the welfare implications of acts even though those implications are not what makes the acts moral or not. Likewise for a desire theorist or an objective-list theorist. The debate in this paper is merely about how best to measure welfare, not about whether or not welfare is the primary component of an ethical theory.

---

[10] For example, Stevenson and Wolfers (2008).

[11] A second set of objections to the evidential defense of revealed preference comes from Sarch (2015). The main objection here concerns the ambiguity in the condition that preferences relevant to welfare consideration be self-interested. van der Deijl (2017) develops a defense of conducting welfare analysis without a complete understanding of what constitutes welfare, an undertaking similar in spirit to that of Hausman and McPherson (2009), though van der Deijl (2017), like Hersch (2015), is critical of Hausman and McPherson's appeal to platitudes (p. 227-229).

[12] "Well-being obviously plays a central role in any moral theory. A theory which said that it just does not matter would be given no credence at all." (Crisp 2017).

III. Behavioral Economics and Preference Purification

I have summarized standard practice of revealed preference welfare economics, some common objections to it, and Hausman and McPherson's evidential defense of it. In this section, I transition to providing a brief overview of behavioral economics and its relationship to welfare evaluations.[13] Then, I will be ready to make my main argument in the following section.

Roughly speaking, behavioral economics 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. Mullainathan and Thaler (2001) offer a definition: "Behavioral economics is the combination of psychology and economics that investigates what happens in markets in which some of the agents display human limitations and complications." Unlike neoclassical economics and the rational choice model, behavioral economics is not a unified theory of behavior; it is a collection of observed "behavioral anomalies" or "behavioral biases" that demonstrate how people deviate from rational choice theory.[14] Researchers have identified several such anomalies and have developed several descriptive theories of behavior to accommodate those anomalies.

One example is the theory of present bias and quasi-hyperbolic discounting. The standard neoclassical model asserts that people make intertemporal decisions using a constant discount rate. For example, when I consider trade-offs between consumption now and consumption one year into the future, I might discount future consumption by 5%. If using a

[13] More thorough overviews and reviews of the field are available in Mullainathan and Thaler (2001), Kahneman (2011), and Thaler (2015).
[14] I will use the terms "behavioral biases" and "behavioral anomalies" interchangeably. Others have used the term "behavioral mistakes," though that term comes with welfare implications that for now I prefer to remain agnostic about.

constant, time-consistent discount rate, then I would also use 5% to discount consumption ten years into the future relative to consumption eleven years into the future. A constant discount rate is time-consistent, since the point in time when I evaluate my options does not affect my preference ranking over options.[15]

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 am evaluating 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. This is not time-consistent if, in ten years (when ten years from now becomes today) the discount rate that I use when evaluating current (in ten years) consumption versus consumption one year in the future (in eleven years) is 10%. Economists can model this type of present bias using quasi-hyperbolic preferences (Laibson 1997), where the discount factor used between any two consecutive future periods (like ten years from now and eleven years from now) is δ, while the discount factor used between now and the following period (like today and next year) is $\beta \times \delta$, where $\beta < 1$ (quasi-hyperbolic discounting

[15] Frederick et al. (2002) describe Paul Samuelson's development of the canonical model of intertemporal choice and its assumption of a constant discount rate. They emphasize "Samuleson's manifest reservations about the normative and descriptive validity of the formulation he had proposed." (p. 351). Samuelson did not endorse it as a normative model, claiming "any connection between utility as discussed here and any welfare concept is disavowed" (Samuelson 1937, p. 161). He also did not even endorse its predictive validity, claiming "it is completely arbitrary to assume that the individual behaves [according to the model]" (Samuelson 1937, p. 159). (Both selections are quoted in Frederick et al. 2002.)

is sometimes referred to as "$\beta\delta$ discounting").[16] Researchers have verified present bias among some subjects in many laboratory experiments and real-world situations.[17]

Consider the problem of making an intertemporal decision over consumption across different periods. Let each period t's consumption level be given by c_t, and the instantaneous utility (that is, the utility gained in period t from consumption in period t) be $u(c_t)$. Period 0 is the present so that c_0 is current-period consumption. The standard, time-consistent model of intertemporal choice asserts that a person acts to as to maximize the net present value of utility discounted with the constant discount factor δ:

$$U = \sum_{t=0}^{T} \delta^t u(c_t)$$

With quasi-hyperbolic ($\beta\delta$) time preferences, the person acts to maximize:

$$U = u(c_0) + \beta \sum_{t=1}^{T} \delta^t u(c_t)$$

A second well-studied behavioral anomaly concerns people's choices over risk. The standard rational choice model assumes that people make risky decisions based on expected utility theory (EUT). If preferences over certain outcomes satisfy some axioms (and so can be called von Neumann-Morgenstern utility functions), then preferences over risky outcomes, like lotteries over payouts, can be described by evaluating the mathematical expectation of the von Neumann-Morgenstern utilities, weighted using the absolute probabilities from the lotteries.

---

[16] A discount factor δ is related to a discount rate r through $\delta = \frac{1}{1+r}$.

[17] For example, DellaVigna and Malmendier (2006), Benhabib et al. (2010), and Montiel Olea et al. (2014). However, Andersen et al. (2014) argue that the evidence for present bias is generally weak, and they find no evidence for $\beta\delta$ discounting.

EUT is elegant and taught to every economics graduate student. However, there are well-known instances where observed preferences are inconsistent with EUT.[18]

An alternative model of choice under risk is prospect theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992). In contrast to EUT, prospect theory posits that people evaluate gains and losses relative to a reference point, rather than evaluate a utility function based on absolute levels of consumption or wealth; that people are loss-averse, meaning they are more averse to a loss relative to the reference point than they are to an equivalent gain; and that people exhibit probability weighting, overweighting small probabilities rather than responding to the objective probabilities of the risky situation. Prospect theory, like present bias, has been verified in many laboratory experiments and real-world situations.[19]

The revealed preference theory of welfare economics and the preference-satisfaction theory of well-being are consistent with a rational choice model of human behavior. There is a well-defined, consistent "welfare function" that maps a person's level of consumption, leisure, and other market and non-market goods and services into his or her well-being. People know how their choices over these goods and services affect their well-being. The rational satisfaction of those preferences maximizes an individual's well-being subject to her constraints.

What are the implications of behavioral economics for welfare economics? That topic is a broad area that I can only briefly summarize.[20] If people's behavior cannot be modeled by rational choice theory, then the link between observed choice and well-being, described in the

[18] One often-cited example is the Allais paradox (Allais 1953), where a person is given two choices, each over two different lotteries. The pair of choices that the majority of people make is inconsistent with any specification of expected utility. However, Conlisk (1989) finds little evidence for this paradox.
[19] For example, Kahneman et al. (1991) and List (2004). But see also Harrison and Swarthout (forthcoming), who find little support for prospect theory from a laboratory experiment and argue instead for a different alternative to expected utility theory: rank-dependent utility.
[20] For more thorough discussions, see Bernheim (2016) or Bernheim and Taubinsky (2018).

previous paragraph and at the heart of the justification for revealed preference welfare analysis, must be called into question.

Much of the economics 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 people behave and the choices they make, but they do not necessarily determine or reveal well-being. Instead, only modified preferences or a subset of those preferences give us information about well-being. One can think of behavioral anomalies as ways in which our preferences deviate from our well-being. Infante et al. (2016) define the preference purification approach thusly: "the task for the planner is to try to reconstruct individuals' underlying or *latent* preferences by simulating what they would have chosen, had they not been subject to reasoning imperfections" (p. 6).

There are several different ways in which one can enact preference purification. Some models have posited that there are two distinct utility functions. The first describes the way that people behave, and has been called "decision utility." The second utility function actually represents well-being, and has been called "experienced utility" (Kahneman et al. 1997).[21] Behavioral anomalies are places where those two utility functions diverge. When conducting welfare analysis, the experienced utility function is what a benevolent planner ought to use to measure welfare, but the planner realizes that individuals act to maximize their decision utility function. If the planner can somehow measure experienced utility, the planner can use policy tools (like taxes or subsidies) to modify the individual's decision utility function so that it coincides with (or at least is closer to) the experienced utility function.

[21] Experienced utility has also been referred to as "true utility" (Gul and Pesendorfer 2007) or "hedonic utility" (Ariely and Norton 2008).

The two behavioral anomalies described earlier (present bias and prospect theory) offer good opportunities for describing how policy analysis and policy design are conducted in the decision utility vs. experienced 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 experienced utility function includes only the long-run discount factor δ. (Thus, O'Donoghue and Rabin (2006) refer to this preference purification approach as the "long-run perspective.") 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 (their experienced utility).[22] The decision utility function is given by

$$U = u(c_0) + \beta \sum_{t=1}^{T} \delta^t u(c_t)$$

However, the experienced utility function, or the individual welfare function, is

$$U = \sum_{t=0}^{T} \delta^t u(c_t)$$

That is, the experienced utility function imposes that the present bias discount factor β be set equal to one. The decision utility function discounts all periods past the present (past c_0) with a lower discount factor. Economists have also used this long-run perspective in the context of environmental policy in Heutel (2015), in the context of Kenyan farmers investing in fertilizer in

[22] "We and other researchers often refer to $\beta < 1$ as representing a "self-control problem" because it reflects a short-term desire or propensity that the person disapproves of at every other moment in her life. Our welfare analysis therefore treats this preference for immediate gratification as an error…" (O'Donoghue and Rabin 2007, p. 1829).

Duflo et al. (2011), and in the context of enrollments in 401(k) retirement accounts in Carroll et al. (2009).

Other papers have conducted welfare analysis using the decision utility vs. experienced utility set-up in the context of prospect theory. The assumption in these papers is that expected utility theory corresponds to experienced utility, while prospect theory corresponds to decision utility. As described by Thaler (2016, p. 1591): "Expected utility theory remains the gold standard for how decisions *should* be made in the face of risk. Prospect theory is meant to be a complement to expected utility theory, which tells us how people *actually* make such choices" (italics added). This interpretation of the normative implications of prospect theory is more controversial than the interpretation of present bias discussed above. It is more acceptable to treat present bias as a mistake or a form of self-control problem than it is to do the same for loss aversion and reference-dependent preferences. It is not so obvious that, if an individual exhibits loss aversion in his decision-making, then this loss aversion ought to be ignored by the social planner in the same way that an individual's present bias β ought to be ignored. Nevertheless, some welfare analyses proceed in this fashion, including Heutel (2019) in the context of energy efficiency policy, Pinto-Prades and Abellan- Perpiñan (2012) in the context of the allocation of health care resources, and Dhami and Al-Nowaihi (2010) in the context of tax evasion. Bleichrodt et al. (2001) is a pioneering use of this approach.[23] Invoking a distinction between decision utility and experienced utility is perhaps the most straightforward preference purification method in behavioral welfare economics, but it is not the only such method.[24]

[23] "In particular, we have argued that loss aversion and probability transformation, two well-documented deviations from expected utility, be recognized and corrected for in utility elicitations." (p. 1509) Infante et al. (2016) summarizes Bleichrodt et al.'s (2001) approach: they "use cumulative prospect theory as the *descriptive* model of choice while retaining expected utility theory as the *normative* model." (p. 8)

[24] For example, Bernheim and Rangel (2009).

I have briefly summarized a small part of the growing literature that uses the preference purification approach in behavioral welfare economics.[25] The following questions remain: how does Hausman and McPherson's evidential defense of welfare economics perform in the context of behavioral economics? Does the evidential defense require preference purification?

IV. The Evidential Defense in Light of Behavioral Economics

Having summarized the evidential defense of welfare economics in section II and the issues related to preference purification and behavioral welfare economics in section III, here I introduce and defend the main claim of this paper: the findings of behavioral economics create problems for the evidential defense of welfare economics.

To begin, it must be stated that the authors of the evidential defense acknowledge this issue. Hausman and McPherson (2009) conclude their essay with an admission that welfare economists must "[correct] for mistakes, biases, and non-self-interested motives that make preferences a misleading guide to welfare" (p. 22-23). These mistakes and biases presumably include those identified by behavioral economics and individuals' deviations from rational choice theory. Furthermore, Hausman (2010) says: "Findings by psychologists and behavioral economists concerning the systematic biases and blunders in people's choices make the understanding of mistakes a pressing problem" (p. 324). It is not that the evidential defense overlooks this issue, but rather that the issue, I will argue, is graver than the authors of the

[25] Though I will not address them here, there are also several arguments that are critical of preference purification. Infante et al. (2016) claim that preference purification presupposes a psychological-normative model where there exists and "inner rational agent… trapped in an outer psychological shell." They argue that this assumption is "fundamentally misconceived;" there is simply no psychological or normative justification for the bifurcation of the inner agent and the outer shell and the assignment of the inner agent's preferences with true, latent preferences. Hausman (2016) offers a rejoinder to their criticism. Finally, Dold (2018) criticizes the preference purification approach from a different perspective, based on Buchanan's notion of "creative choice."

evidential defense acknowledge, and it deserves more than a passing caveat (or, as Hausman (2010) states, "a note of caution [that] should be marked fortissimo" (p. 341)).

I proceed by considering how one would attempt to use the evidential defense to justify the use of revealed preference welfare economics in a world in which 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 often inconstant and predictably irrational, in accordance with various models from behavioral economics. Can I still use the evidential defense to justify conducting welfare economics, i.e., to justify using revealed preference to measure well-being? 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*. Clearly, some mistaken preferences cannot be good evidence for well-being – if a tourist to England prefers to drive his car on the right side of the road believing it to be safe and legal, a welfare economist ought not to infer that he enhances his well-being through a head-on collision. Mistaken preferences and non-self-interested preferences ought to be laundered from the

preferences that can be used in welfare analysis, an algorithm that is similar in spirit to the preference purification used in behavioral welfare economics. 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. 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 fairly 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.

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. In practice, this means that a welfare economist makes no attempt to disentangle or eliminate preferences that do not conform to rational choice theory and may be inconsistent, 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.[26] 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

[26] 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).

– nowhere am I mistaken 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 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 simply the day of the week that I am doing the evaluating. (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.)

The evidential defense does not require or claim that preferences give *perfect* evidence for what promotes well-being, but merely that preferences give *the best available* evidence for it. A reply to the argument made in the preceding paragraph is to admit that the inconsistencies are clearly not *constitutive* of well-being but are nonetheless the best available evidence for it. That is, the best available evidence on Sunday comes from the preferences that I exhibit on Sunday, while the best available evidence on Tuesday comes from the preferences that I exhibit on Tuesday, even though both sets of preferences cannot be perfect evidence for what constitutes well-being. The time-inconsistency of this evidence for what promotes well-being is unsatisfying, but according to this argument is something that the applied welfare economist will have to deal with if she wants to use the best-available evidence.

I find this reply unconvincing. The inconsistent pieces of evidence for what promotes well-being are only the best available evidence if one insists on not purifying the preferences of their time-inconsistency. It is much more reasonable to conclude that even better evidence for what promotes well-being is available – namely, the evidence from the preferences, coupled with some knowledge about features of human decision-making like procrastination. One requires neither a theory of well-being nor a sophisticated theory of psychology and decision-making to conclude that people often procrastinate and do so in reasonably predictable ways, and that procrastination systematically leads to preferences and choices not doing a good job indicating what promotes well-being. To refuse to purify preferences is to claim that preferences that are clearly formed through procrastination (like preferring to not go to the gym on Tuesday even though earlier in the week you had planned to go on Tuesday) are the best evidence we have for what constitutes well-being, which amounts to ignoring facts about human decision-making that are reasonably uncontroversial.

For a second example of how not purifying preferences leads to unsatisfying conclusions for the evidential defense, consider the well-known behavioral anomaly present in a cafeteria: my preferences between cake and vegetables depend on their relative placement in the cafeteria line. If cake is presented more closely to me (e.g. at eye-level), then I am more likely to prefer and choose cake. If employing the evidential defense without purifying preferences of this behavioral anomaly, we conclude that the best evidence for which food promotes my well-being depends on the placement of the food in the cafeteria. This, too, is unsatisfying. It seems implausible that what in fact most improves my well-being depends on its location in the cafeteria, so it is unsatisfying to claim that the best evidence is so dependent on location.

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, according to behavioral economics. 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 to me that the defenders of the evidential defense would endorse this interpretation of it. Indeed, as I have previously quoted, Hausman and McPherson (2009) admit that welfare economists must correct for "mistakes, biases, and non-self-interested motives."

A reply to my argument could be offered in the spirit of the "opportunity criterion" of welfare analysis developed in Sugden (2004).[27] Sugden argues that opportunity (consumer sovereignty), in and of itself, contributes to people's well-being, not merely because opportunity allows people to satisfy coherent preferences and maximize some coherent welfare function. He shows that under certain assumptions, a free market will achieve the opportunity criterion, which is analogous to the concept of Pareto efficiency, though defined over opportunities rather than preference satisfaction. According to this criterion, inconsistent preferences like the examples I have just described are not problematic for a welfare economist, since the goal is not to maximize a coherent welfare function. Rather, the goal is (roughly) to maximize opportunity, and if people happen to use those opportunities in ways that seem irrational or inconsistent (e.g.

[27] See also Sugden (2009).

subject to framing or procrastination), that is not ethically problematic. A devotee of the opportunity criterion would likely not be as bothered by the above-mentioned inconsistencies.[28]

The appeal to the opportunity criterion is not an adequate reply to my argument that failing to purify preferences of behavioral anomalies leads to unsatisfying and implausible conclusions for the evidential defense. The evidential defense claims that revealed preference provides the best evidence for what promotes well-being regardless of the theory of well-being, and therefore claims that even under an opportunity criterion theory of well-being, revealed preference provides the best evidence. The reason that devotees of the opportunity criterion are not bothered by preference reversals and similar inconsistencies is that they aren't interested in revealed preference in the first place. They aren't looking at preferences to measure welfare; they want to look at opportunity sets to measure it. How this is done in practice is unclear, but it is definitely not through standard revealed preference economics. Advocates of the opportunity criterion have no interest in defending the practice of welfare economics via the evidential defense or via any other defense. It is thus inappropriate to use the opportunity criterion as a reply to my argument about the problems with the evidential defense when not purifying preferences. However, this discussion hints at the issue that the evidential defense cannot truly be sustained without at least some form of a theory of well-being, an issue I will return to in the following section.

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. This strategy is likely what McPherson and Hausman

[28] "Any increase in [an] individual's lifetime opportunity is good for her in an unambiguous sense… This is true whether or not her actions across time are consistent with any one set of coherent preferences." (Sugden 2004, p. 1018)

have in mind in their asides mentioning the caveats of their analysis in the face of behavioral economics. To be more clear 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. This third condition is added to purify preferences of the behavioral anomalies (perhaps now it is appropriate to refer to them as "behavioral mistakes") before making welfare conclusions from them.

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 begs 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 and loss aversion. However, the justification for purifying these types of preferences is not immediately clear. A key feature of the evidential defense, according to its defenders, is that it is agnostic about what constitutes welfare; it does not rely on or endorse any particular theory of welfare.[29] But, how can we justify purifying preferences of their behavioral anomalies without relying on

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

some theory of welfare? (For example, a theory that claims that what constitutes welfare cannot be time-inconsistent in the way that preferences sometimes are.) 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. Why not? If we do not purify preferences of these behavioral anomalies, then the evidential defense is on shaky ground, as I have argued earlier. But that fact is not an argument for purifying them or a justification for the evidential defense.

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. This critique of the evidential defense is reminiscent of some discussion in Hersch (2015), who, as described earlier, questions Hausman and McPherson's (2009) reliance on platitudes for justifying the evidential defense and the claim that it does not rely on a theory of well-being. The crux of Hersch's argument is 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. If there are platitudes

that can be relied on that relate to behavioral anomalies like time inconsistency, loss aversion, or framing, I am not aware of any.

Hausman and McPherson give the example of preferences over the existence of endangered species like Siberian tigers and whooping cranes. 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, 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. 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. Admittedly, in the case of procrastination caused by time-inconsistent preferences, it does appear that a reasonably uncontroversial platitude might be able to be used to identify when to purify preferences. But, I doubt that this is generally true of most behavioral anomalies. Consider loss aversion. When a value function exhibits a kink at the reference point (as is characteristic of loss aversion), I know of no platitude that can be applied to claim that this kink is welfare-irrelevant.

Let us for now ignore the difficulties justifying the purification of preferences that exhibit behavioral anomalies, and move on to the second question that their proposed purification begs: how exactly would we purify these preferences? As with the first question, when addressing this question, 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 our workhorse example of 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. This is the long-run criterion, which amounts to asserting that the decision utility function includes present-biased, $\beta\delta$ time preferences but that the experienced utility function features time-consistent preferences (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 based 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 loss aversion. The closest thing here to a standard methodology of preference purification is to assert that the decision utility function is characterized by loss aversion and prospect theory but that the experienced utility function is characterized by expected utility theory (Bleichrodt et al.

2011). 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). Admittedly, platitudes may advise us that framing can create biases that should not count towards welfare analysis, or that two equivalent decisions that differ only in their presentation or framing should not lead to contradictory welfare evaluations. But, the jump from these broad, general platitudes to very specific definitions of decision and experienced utility functions is a large one.

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. That is, according to this counterargument, the welfare economist doesn't require details on how the two discounting parameters (β and δ) relate to welfare, or details on how the value function and the probability weighting in prospect theory relate to welfare, or exactly how we mathematically manipulate the evidence from preferences and choices to measure well-being. If this is true, and the welfare economist merely needs a rough way to purify preferences, 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 time-inconsistent 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.

My argument in this section is similar in spirit to that of Hersch (2015), described earlier, which also criticizes the evidential defense. Hersch's main critique is that the appeal to platitudes fails and does not uniquely defend preferences as the best evidence of what promotes

well-being. The critique that I present in this section does not rely on this same attack. My critique is not that platitudes are not the best evidence for what promotes well-being, but in fact that there are no platitudes that can be used to purify preferences of their behavioral anomalies.

The appeal to platitudes invoked by Hausman and McPherson (2009) does not reply to my objections here. While the invocation of platitudes may work in determining that preferences that are misinformed or non-self-interested should not be used as evidence for what promotes well-being, the invocation of platitudes does not work for purifying preferences of behavioral anomalies. Hausman and McPherson (2009) claim "the mistakes individuals make about their own good will be to some extent unsystematic" (p. 16). My claim here is that extent to which the mistakes are *unsystematic* is smaller than Hausman and McPherson may have been willing to admit, given the extent of behavioral anomalies. Behavioral economics has identified systematic mistakes that people often make, and these cannot be erased from welfare consideration merely by platitudes.

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 with preferences that are purified of behavioral anomalies, then this purification cannot be done without some theory of well-being more thorough than platitudes about what does and what doesn't count towards one's well-being. The fact that behavioral anomalies are systematic and predictable is important; if errors were symmetric or unsystematic then the evidential defense would not be as troubled by them.

Where does this leave the applied welfare economist? In the concluding section, I will describe how he or she can proceed.

V. Modifying the Evidential Defense to Accommodate Behavioral Anomalies

In light of these issues with the evidential defense, we have several possible responses. First, we could abandon the evidential defense and abandon any hope of justifying welfare economics. Second, we could fall back on the preference-satisfaction theory of well-being – make explicit the implicit assumption behind much practiced welfare economics that preference satisfaction is constitutive of well-being. Then we would not need to worry about any of the aforementioned issues with the evidential defense or with the evidential defense at all. Both of these options are rather extreme and unwarranted.

A third option is more moderate and tenable. We could assert something like the purified preference-satisfaction theory of well-being – that preferences are indeed constitutive of welfare, but only when those preferences are purified of their behavioral anomalies. This would be a modification of one of the "laundered" preference theories of well-being, with the modification being that the laundering (or purifying) is ridding the preferences of their behavioral biases. Such a theory, of course, would be subject to many of the same critiques of other preference satisfaction theories of well-being including others with laundered preferences, as I summarized earlier in this paper.

The fourth option will be the topic of this last section of this paper. The fourth option is to modify the evidential defense of welfare economics in such a way as to allow it to accommodate behavioral anomalies. As described above, the evidential defense encounters

problems due to behavioral anomalies – it is not plausible to disregard the behavioral anomalies when conducting welfare analyses and it is not justified to purify preferences of their behavioral anomalies without a theory of well-being, as the evidential defense strives to omit. My claim here will basically be that the evidential defense can 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.

I claim that the evidential defense of welfare economics should be modified to assert that some common preferences and behaviors of people are clearly not preferences and behaviors that act so as to maximize their well-being. This can be true even of preferences that are self-interested and well-informed (and that thus satisfy the more moderate purification called for by the supporters of the original evidential defense). People can exhibit self-control problems, procrastination, distractions by irrelevant factors, and other features that render their preferences poor guides to their well-being. By "poor guides," I do not mean small mistakes here and there, but potentially large deviations between preferences and well-being (think not of the occasional procrastinator but instead of the heroin addict). These types of preferences cannot be evidence for well-being, and they should not be used in welfare analysis. I would thus add 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."

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. For example, preferences that exhibit self-control issues, procrastination, or distraction by irrelevant information (like the placement of food items in a cafeteria) are not good evidence of well-being because choices made with those preferences cannot be choices that promote well-being, or at least cannot be choices that systematically and reliably promote well-being. It cannot be the case that what makes you better off depends on where it is placed in the cafeteria; it cannot be the case that what makes you better off depends on the day of the week that you are making your plans; it cannot be the case that what makes you better off depends on whether it is framed as a gain or a loss. These statements appear to me to be uncontroversial, nevertheless to make these statements one needs some form or outline of a theory of well-being that rules out (uncontroversially) some outcomes that cannot constitute well-being. This outline of a theory of well-being 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.[30]

Given this outline of a theory of well-being, how does the modified evidential defense operate? It need not go so far as to dictate the form of the experienced utility function, or to dictate exactly how preferences are to be purified (e.g., the theory need not claim that well-being is defined by discounting only using the long-run discount factor δ and not the present bias discount factor β). But, it can still defend the *practice* of using a particular experienced utility function on the grounds that the welfare evaluated with that experienced utility function is the best available evidence for well-being that welfare economists have. In other words, while the

[30] I do not need to provide a comprehensive list of cases where preferences are not giving evidence for what does not promote well-being, and I do not claim that the cases listed in this paragraph are exhaustive. Like the Bernheim and Rangel (2009) approach, which allows for ancillary conditions that are welfare-irrelevant but does not list all such ancillary conditions, here I merely provide a framework for modifying the evidential defense.

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 an experienced utility function, which is estimated from choices but not perfectly corresponding to them the way that the decision utility function is, is the best available evidence of well-being though it does not define or constitute well-being.

In arguing for such a defense of behavioral welfare economics (indeed, this is a defense of the standard preference purification method of behavioral welfare economics, just as the original evidential defense is a defense of the standard method of neoclassical welfare economics), we need not be wedded to any particular experienced 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 a number of ways to purify preferences of their behavioral anomalies. For the example of prospect theory, one might argue that probability weighting represents a behavioral bias that needs to be purified, but that loss aversion is not a bias and is indeed truly reflective of welfare. This would be an alternative to the purification approach of Bleichrodt et al. (2001) and others, who strip the decision utility function of both its probability weighting and its loss aversion before defining the experienced utility function. The modified evidential defense need not pin down a particular purification approach. 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 the best 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.

In conclusion, I have summarized Hausman and McPherson's (2009) evidential defense of welfare economics, and I have summarized behavioral welfare economics and the preference purification approach. My central thesis in this paper has been that the findings of behavioral economics present difficulties for the evidential defense that its originators did not sufficiently account for. It is implausible to ignore behavioral anomalies when using the evidential defense, and it is impossible to purify preferences of their behavioral anomalies without imposing some rudimentary theory of well-being, as the evidential defense's originators claim can be done. I propose that a modified evidential defense of welfare economics can survive, one in which one must rely on a rudimentary and uncontroversial theory of well-being in which some behaviors that cannot plausibly be said to promote well-being are eliminated. With this modified defense, behavioral welfare economists have an evidential justification for their approach.

References

Allais, M. (1953). "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine". *Econometrica.* 21 (4): 503–546.

Allcott, Hunt. "Social norms and energy conservation." *Journal of Public Economics* 95, no. 9-10 (2011): 1082-1095.

Allcott, Hunt, and Judd B. Kessler. "The welfare effects of nudges: A case study of energy use social comparisons." *American Economic Journal: Applied Economics*, 11, no. 1 (2019): 236-76.

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.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. "Discounting behavior: A reconsideration." *European Economic Review* 71 (2014): 15-33.

Angner, Erik. "Well-being and economics." *The Routledge Handbook of the Philosophy of Well-Being (London: Routledge, 2015)* (2015): 15-1.

Ariely, Dan, and Michael I. Norton. "How actions create–not just reveal–preferences." *Trends in cognitive sciences* 12, no. 1 (2008): 13-16.

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.

Bergstrom, Ted. "Benefit cost analysis and the entanglements of love." Working paper (2003).

Bernheim, B. Douglas. "The good, the bad, and the ugly: A unified approach to behavioral welfare economics." *Journal of Benefit-Cost Analysis* 7, no. 1 (2016): 12-68.

Bernheim, B. Douglas, and Antonio Rangel. "Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics." *The Quarterly Journal of Economics* 124, no. 1 (2009): 51-104.

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.

Bleichrodt, Han, Jose Luis Pinto, and Peter P. Wakker. "Making descriptive use of prospect theory to improve the prescriptive use of expected utility." *Management science* 47, no. 11 (2001): 1498-1514.

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.

Conlisk, John. "Three variants on the Allais example." *The American Economic Review* (1989): 392-407.

Crisp, Roger, "Well-Being", *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2017/entries/well-being/.

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

Dhami, Sanjit, and Ali Al-Nowaihi. "Optimal taxation in the presence of tax evasion: Expected utility versus prospect theory." *Journal of Economic Behavior & Organization* 75, no. 2 (2010): 313-337.

Dold, Malte F. "Back to Buchanan? Explorations of welfare and subjectivism in behavioral economics." *Journal of Economic Methodology* 25, no. 2 (2018): 160-178.

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.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. "Time discounting and time preference: A critical review." *Journal of Economic Literature* 40, no. 2 (2002): 351-401.

Gul, Faruk, and Wolfgang Pesendorfer. "Welfare without happiness." *American Economic Review* 97, no. 2 (2007): 471-476.

Harrison, Glenn W., and Jia Min Ng. "Evaluating the expected welfare gain from insurance." *Journal of Risk and Insurance* 83, no. 1 (2016): 91-120.

Harrison, Glenn W., and J. Todd Swarthout. "Cumulative prospect theory in the laboratory: A reconsideration." Forthcoming in G.W. Harrison and D. Ross (eds.), *Models of Risk Preferences: Descriptive and Normative Challenges* (Bingley, UK: Emerald, Research in Experimental Economics).

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

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.

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

Heutel, Garth. "Prospect theory and energy efficiency." *Journal of Environmental Economics and Management* 96 (2019): 236-254.

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.

Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

Kahneman, Daniel, and Amos Tversky. "Prospect theory: An analysis of decision under risk." *Econometrica: Journal of the Econometric Society* (1979): 263-291.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. "Anomalies: The endowment effect, loss aversion, and status quo bias." *Journal of Economic Perspectives* 5, no. 1 (1991): 193-206.

Kahneman, Daniel, Peter P. Wakker, and Rakesh Sarin. "Back to Bentham? Explorations of experienced utility." *The Quarterly Journal of Economics* 112, no. 2 (1997): 375-406.

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

List, John A. "Neoclassical theory versus prospect theory: Evidence from the marketplace." *Econometrica* 72, no. 2 (2004): 615-625.

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.

Mullainathan, Sendhil, and Richard H. Thaler. *Behavioral economics*. No. w7948. National Bureau of Economic Research, 2000.

Nussbaum, Martha. "Capabilities as fundamental entitlements: Sen and social justice." *Feminist Economics* 9, no. 2-3 (2003): 33-59.

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.

Reiss, Julian. *Philosophy of economics: A contemporary introduction*. Routledge, 2013.

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 K. "Choice functions and revealed preference." *The Review of Economic Studies* 38, no. 3 (1971): 307-317.

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

Stevenson, Betsey, and Justin Wolfers. "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox." *Brookings Papers on Economic Activity* (2008).

Sugden, Robert. "The opportunity criterion: consumer sovereignty without the assumption of coherent preferences." *American Economic Review* 94, no. 4 (2004): 1014-1033.

Sugden, Robert. "Market simulation and the provision of public goods: a non-paternalistic response to anomalies in environmental evaluation." *Journal of Environmental Economics and Management* 57, no. 1 (2009): 87-103.

Thaler, Richard H.. *Misbehaving: The making of behavioral economics*. New York, NY: WW Norton, 2015.

Tversky, Amos, and Daniel Kahneman. "Advances in prospect theory: Cumulative representation of uncertainty." *Journal of Risk and Uncertainty* 5, no. 4 (1992): 297-323.

van der Deijl, Willem. "Are measures of well-being philosophically adequate?." *Philosophy of the Social Sciences* 47, no. 3 (2017): 209-234.