NBER WORKING PAPER SERIES

BEHAVIORAL WELFARE ECONOMICS

B. Douglas Bernheim

Working Paper 14622
http://www.nber.org/papers/w14622

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2008

This paper was prepared for the 2008 EEA-ESEM Meetings in Milan, Italy, where it was delivered as the Schumpeter Lecture. My thinking on this topic was profoundly shaped by numerous conversations with Antonio Rangel, as well as by our joint work, summarized in Section 4. I am indebted to an anonymous referee for thoughtful and insightful comments. I gratefully acknowledges financial support from the NSF (SES-0452300 and SES-0752854). The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Behavioral Welfare Economics
B. Douglas Bernheim
NBER Working Paper No. 14622
December 2008
JEL No. D01,D60,H40

ABSTRACT

This paper discusses several competing proposals for general normative frameworks that would encompass non-standard models of choice. Most existing proposals equate welfare with well-being. Some assume that well-being flows from the achievement of well-defined objectives, and that those objectives also guide choices; the trick is to formulate a framework in which less-than-completely coherent choice patterns reveal the unobserved objectives. Others are predicated on the contention that well-being, and hence welfare, is directly measurable. Both of those approaches encounter serious conceptual difficulties. An alternative approach, developed by Bernheim and Rangel [2009], defines welfare directly in terms of choice. It entails a generalized welfare criterion that respects choice directly, without requiring any rationalization involving potentially unverifiable assumptions concerning underlying objectives and their relationships to choice. Because useful behavioral theories generally envision a substantial degree of underlying coherence in behavior, that criterion leads to a rich and tractable normative framework.

B. Douglas Bernheim
Department of Economics
Stanford University
Stanford, CA 94305-6072
and NBER
bernheim@stanford.edu

1 Introduction

Positive theories of decision making from the field behavioral economics are steadily infiltrating mainstream thought. As they find their way into policy analysis, their use inevitably raises perplexing questions concerning welfare. In particular, if an individual's choices are neither fully rational nor completely coherent, how can they serve as the foundation for a compelling welfare standard? Lacking a generally accepted normative framework, economists have tended to rely on *ad hoc* criteria for particular positive models, offering justifications based on loose and inevitably controversial intuition. Plainly, that state of affairs is unsatisfactory.

Recent work lays out a number of competing proposals for general normative frameworks that would encompass non-standard models of choice. Each proposal begins with an implicit or explicit definition of welfare. Most existing proposals equate welfare with well-being. Some assume that well-being flows from the achievement of well-defined objectives and that those objectives also guide choices; the trick is to formulate a framework in which less-than-completely coherent choice patterns reveal the unobserved objectives (possibly in combination with other types of evidence). Others are predicated on the contention that well-being, and hence welfare, is directly measurable. Though the notion of welfare as well-being is intuitively appealing, both approaches – revealed well-being and measured well-being – encounter serious conceptual difficulties, which I discuss in Sections 2 (on revealed well-being) and 3 (on measured well-being).

In a recent paper (Bernheim and Rangel [2009]), Antonio Rangel and I proposed an alternative normative framework that defines welfare directly in terms of choice rather than well-being or underlying objectives (see also Bernheim and Rangel [2007, 2008]). That perspective has a long history dating back to the origins of the revealed preference revolution, and it is consistent with the choice-theoretic foundations of standard welfare analysis. When the Weak Axiom of Revealed Preference is satisfied, the statement "x is strictly revealed preferred to y" (xPy) is equivalent to the observation that x (and not y) is chosen from the set

$\{x, y\}$. Thus, one can determine whether xPy directly from choice patterns without relying on any underlying rationalization. Furthermore, one does not require a rationalization to justify normative judgments. Arguably, choices provide appropriate guidance either because they are good proxies for well-being, because a general policy of deference to choice tends to improve well being by promoting better governance, or simply because choice per se is normatively compelling. Accordingly, we propose a generalized welfare criterion that respects choice directly, without requiring any rationalization involving potentially unverifiable assumptions concerning underlying objectives and their relationships to choice. Because useful behavioral theories generally envision a substantial degree of underlying coherence in choice, that criterion leads to a rich and tractable normative framework. I explain and discuss our framework in Section 4.

The literature also contains proposals that envision more radical departures from the conventional framework. They define welfare in terms of concepts such as opportunity, functionality, and political sustainability. Because this paper is not intended as a comprehensive survey of behavioral welfare economics, I mention those other concepts only briefly in Section 5 along with concluding comments.

2 Welfare as revealed well-being

According to one interpretation, standard normative analysis evaluates the decision maker's well-being according to her true objectives, which her choices reveal; see, e.g., Sen [1973] or, more recently, Koszegi and Rabin [2008a]. Attempts to generalize the standard framework that embrace this interpretation encounter serious conceptual difficulties. The central problem is easily stated: because the behaviors of interest by definition defy conventional rationalizations, one must open the door to unconventional rationalizations. But as a general matter one can offer many unconventional rationalizations for any particular behavioral pattern (even when behavior satisfies standard axioms). Thus, knowledge of a choice correspondence may shed insufficient light on objectives, and hence on the mapping from the objects of choice to well-being. One can attempt to identify welfare either partially or completely by restricting the set of allowable unconventional rationalizations, but useful restrictions are difficult to justify.

It is also worth emphasizing that this approach presupposes the existence of well-behaved (if exotic) preferences. Some psychologists have disputed the validity of that premise. Algorithmic decision processes, including ones that are "procedurally rational" (in the sense of Simon [1976]), can generate systematic choice patterns even without coherent underlying objectives.[1] For such models of behavior, a welfare framework that requires the identification of "true" preferences is vacuous.

The literature examines two distinct strategies for rationalizing non-standard choice patterns: broaden the preference domain while maintaining the assumption that choice always maximizes a single coherent objective function, or relax the latter assumption, either by adopting a model that accounts for divergences between preference and behavior, or by supposing that the individual pursues multiple conflicting objectives.

2.1 Broadening the preference domain

To someone who equates welfare with "true" well-being, standard welfare analysis begins with the assumption that the individual is endowed with an objective function $u(x)$ that depends only on the identity of the chosen object, x (see, e.g., Koszegi and Rabin [2008a]). In principle, well-being could also depend on other aspects of the decision problem. Indeed, if we wish to account for behavior that violates standard choice axioms, and if we insist on maintaining the assumption that choices always maximize a well-behaved objective function, we are essentially compelled to broaden the preference domain, for example by treating the sets from which objects are chosen as arguments of the utility function (as implied by Gul and Pesendorfer's [2001] axiomitization of temptation). In a penetrating essay, Koszegi and Rabin [2008a] asked whether information about choices permits us to recover sufficient

[1] For example, as Mandler, Manzini, and Mariotti [2008] demonstrate, the behavior of an individual who makes a selection by applying a checklist satisfies WARP, and consequently can be represented as maximizing a utility function.

information concerning well-being to perform meaningful welfare analysis once we allow for more flexible objective functions. This section summarizes and elaborates on their central insights.

The Framework. Let's begin with some notation. Let $S = (X, B, F, d)$ be a decision problem; it consists of a set of objects $X \subset \mathbb{X}$ (where $\mathbb{X}$ is the universe of possible alternatives), behaviors B (strategies for choosing objects), a function F mapping each behavior $b \in B$ into an object $F(b) \in X$, and conditions d (such as anchors, style of information presentation, and so forth).[2] So, for example, S might require an individual to select from three objects, x_1, x_2, and x_3, by sequentially deleting two after observing a light flash either red or green. In that case, $X = \{x_1, x_2, x_3\}$, B consists of all ordered pairs (x_j, x_k) with $j \neq k$ (interpreted as indicating that the individual strikes x_j and then x_k), $F(x_j, x_k) = x_i$ where $i \notin \{j, k\}$, and $d = \{\text{red, green}\}$. Let $\beta(S)$ denote the behavior chosen for S. To keep the discussion free from distracting technicalities, I will generally write as if β is single-valued; the detail-conscious reader can easily reformulate the arguments to accommodate correspondences. Let $\chi(S)$ denote the object chosen for the decision problem S; that is, $\chi(S) = F(\beta(S))$.

We will broaden the preference domain by writing well-being as a function of the form $u(b, S)$. There is no need to include the chosen object x, as it is uniquely implied by the mapping F, given b. By writing well-being as $u(b, S)$ rather than $u(x, S)$, I allow for slightly greater generality, in that well-being can depend on the manner in which an object is selected. For instance, continuing the example from the last paragraph, the decision maker might feel happier choosing x_1 by eliminating x_2 and then x_3, rather than by eliminating x_3 and then x_2. To minimize technicalities, I will generally write as if the behavior b that maximizes $u(b, S)$ is unique; once again the detail-conscious reader can generalize the arguments to correspondences.

[2] This definition differs from Koszegi and Rabin's, in that it makes various components of the decision problem explicit rather than implicit. I choose a more explicit definition because it facilitates greater expositional precision and clarity.

Obviously, objectives of the form $u(b, S)$ can account for a wide variety of choice anomalies. Take the case of choice reversals, which occur when there are two decision problems, S_1 and S_2, such that$\chi(S_1) \neq \chi(S_2)$ but $\chi(S_i) \in S_j$ for $i \neq j$ (so that $\chi(S_1)$ is chosen in S_1 even though $\chi(S_2)$ is available, and $\chi(S_2)$ is chosen in S_2 even though $\chi(S_1)$ is available). Such behavior does not maximize any objective function of the form $u(x)$ defined over the objects in X. We can potentially resolve the apparent inconsistency by broadening the preference domain to include behaviors and/or features of the decision problem. Choice data tell us only that $u(\beta(S_1), S_1) > u(b, S_1)$ for $b \notin \beta(S_1)$ in B_1 (including behaviors for which $F(b) = \chi(S_2)$) and $u(\beta(S_2), S_2) > u(b, S_2)$ for $b \notin \beta(S_2)$ in B_2 (including behaviors for which $F(b) = \chi(S_1)$); no inconsistency remains.

The identification problem. We would like to know whether choice patterns identify $u(b, S)$. Without further restrictions on the objective function, identification is plainly hopeless. Consider the following simple illustration: a decision maker must choose between a "fair" allocation that divides a monetary prize equally with a partner, and an "unfair" allocation in which the decision maker receives the entire prize. Suppose he chooses the fair allocation to avoid guilt. In that case, he might nevertheless experience greater well-being if someone else imposed the unfair allocation, leaving him guiltless. We might hope to detect such a preference by creating a decision task in which he chooses between making the choice himself and having someone else impose the unfair allocation (a meta-choice). However, even in that setting, the decision maker remains fully responsible for the outcome, and therefore may still choose the fair allocation to avoid guilt. Even if he chooses the fair allocation in every conceivable decision problem that offers both the fair and unfair allocations, he may still experience greater well-being when someone else imposes the unfair allocation. Consequently, a planner acting on his behalf cannot determine which allocation will provide him with greater well-being.

Without further restrictions, it is plainly impossible to glean any useful information concerning $u(b, S)$. For any decision problem S (including ones with singleton opportunity sets), choice patterns tell us that $u(\beta(S), S) > u(b, S)$ for all $b \neq \beta(S)$ in B. But that observation does not allow us to compare $u(\beta(S_1), S_1)$ and $u(\beta(S_2), S_2)$; consequently we cannot determine which decision problem makes the individual better off.[3] Intuitively, we might hope to resolve that issue by providing him with a meta-choice, S^m, between S_1 and S_2. However, from the meta-choice, we learn only that $u(\beta(S^m), S^m) > u(b, S^m)$ for all $b \neq \beta(S^m)$ in B^m; we obtain no information concerning $u(\beta(S_1), S_1)$ or $u(\beta(S_2), S_2)$.

Notably, our conclusion does not depend on the properties of the choice mapping β. Once we allow for the possibility that well-being depends on the characteristics of a decision problem, we simply cannot say whether any given change in the individual's environment makes him better or worse off, irrespective of whether choice patterns satisfy standard axioms. To illustrate, suppose the well-being function is of the following form (inspired by Gul and Pesendorfer's model of temptation):

$$u(b, S) = f(F(b)) - \alpha[\max_{x \in X} g(x) - g(F(b))], \tag{1}$$

where g and f are scalar-valued functions and $\alpha > 0$. Note that well-being depends only on the chosen object and the set of available objects. We interpret $g(x)$ as the virtuousness of alternative x; thus, $\max_{x \in X} g(x) - g(F(b))$ represents the guilt the decision maker experiences when he fails to pick the most virtuous alternative. For simplicity, assume X is a finite set and that α is very large. In that case, the decision maker always maximizes g, so standard choice axioms are satisfied. Moreover, substituting the expression $g(F(\beta(S))) = \max_{x \in X} g(x)$ into equation (1), we see that for any problem S, the level of well-being associated with his chosen behavior is $u(\beta(S), S) = f(\chi(S))$. In other words, because he always acts to eliminate guilt, guilt-avoidance drives his choices (revealing g), but has no bearing on his well-being. Instead, the level of well-being achieved in any decision problem is governed by f, which has no influence over choice. Any two well-being func-

[3] Neither can we rank unchosen alternatives within a decision problem. That limitation is of no consequence, however, because there is no environment in which the individual would end up with an unchosen alternative within the opportunity set from which he chooses. In particular, a policy that compels him to select an otherwise unchosen alternative changes the opportunity set.

tions that differ only in f and that respect fixed upper and lower bounds are observationally equivalent in terms of their implications for choice. Therefore, well-being is completely unidentified.

Potential identifying restrictions. We could, of course, impose identifying restrictions on the well-being function. For example, we might assume that well-being is invariant with respect to the highest level characteristics of sufficiently high level meta-problems.[4] Formally, the assumption would imply that, for any such meta-problem S^m, we have $u([S_k, b], S^m) = u(b, S_k)$ (where the notation $[S_k, b]$ implies that the individual begins by selecting whichever actions deliver the decision problem S_k, and then selects the behavior b in S_k). In that case, the individual will choose either $[S_1, \beta(S_1)]$ or $[S_2, \beta(S_2)]$, and his meta-choice reveals whether $u(\beta(S_1), S_1)$ is greater or less than $u(\beta(S_2), S_2)$.

The identifying assumption proposed in the previous paragraph is of course not testable based on choice patterns. It may nevertheless strike the reader as appealing, in that it seems to rule out complex and therefore potentially implausible models of well-being. Unfortunately, that appearance is deceiving. For example, one can show that the assumption bars the natural possibility that well-being depends on the chosen object, the set of available objects, and nothing else (i.e., any function for the form $u(b, S) = U(F(b), X)$, where U is sensitive to X, as in equation (1)). For well-being functions belonging to that class, meta-choices are unenlightening: a meta-choice that begins with a selection from the meta-set $\{X_1, ..., X_K\}$ provides precisely the same information concerning well-being as a simple choice from the set $X = \cup_{k=1}^{K} X_k$.

In some instances, we might also hope to identify well-being through a contingent assumption: if the highest level choices in all level k meta-problems reflect well-behaved preferences over potential continuation problems *and nothing else*, then we assume that well-being depends only on the selected level $k-1$ meta-problem and the choice made within that problem.

[4] We define the level of a meta-problem recursively. A choice among elements of $\mathbb{X}$ is a level 0 meta-problem; a choice among meta-problems of level $z-1$ or less (with at least one of level $z-1$) is a level z meta-problem.

Among other things, that assumption justifies conventional welfare economics provided that standard choice axioms are satisfied. It may strike the reader as potentially reasonable and innocuous. However, as a formal matter, it is difficult to justify. Generally speaking, it rules out separable well-being functions of the form $u(b, S) = v(b) + w(S)$, including the one shown in equation (1).

Other potential identifying restrictions pertain to the role of the planner. Let's imagine that the planner must provide the individual with one of two decision problems, S_1 or S_2. (Either or both of those problems may be degenerate; e.g., the planner may simply choose the object.) When the planner chooses S_k, the individual faces the decision problem S_k^P (where the P indicates that the decision problem arose from a broader setting in which the planner made particular choices). To advise the planner, we must determine the sign of $u(\beta(S_1^P), S_1^P) - u(\beta(S_2^P), S_2^P)$. So far, I have proceeded as if our task is to identify the sign of $u(\beta(S_1), S_1) - u(\beta(S_2), S_2)$. The relevance of that task is apparent only if we expect to formulate our advice based on the assumption that the individual views S_k^P and S_k as identical (i.e., $u(b, S_k^P) = u(b, S_k)$). Certainly, it seems natural to assume that the individual's well-being is unaffected by the planner's consideration and rejection of another alternative, but as we have seen, that assumption does not resolve our difficulties, as the identification of $u(\beta(S_1), S_1)$ and $u(\beta(S_2), S_2)$ is problematic. There is another possibility worth considering: assume the individual derives the same well-being from the planner's choice and from an equivalent private choice. According to that assumption, there is a decision problem S^I in which the individual chooses between S_1 and S_2 for which $u(b, S_k^P) = u((S_k, b), S^I)$; that is, he derives the same pleasure from an outcome regardless of whether he or the planner selects S_k, given the same alternatives. Let's assume without loss of generality that the individual chooses S_1 in S^I. In that case, the fact that $u(\beta(S^I), S^I) > u(b, S^I)$ for all $b \in B^I$ tells us that $u(\beta(S_1^P), S_1^P) > u(b, S_k^P)$ for all $(b, S_k^P) \neq (\beta(S_1^P), S_1^P)$. Without additional identifying restrictions, we can learn nothing more about the function $u(b, S^I)$; however, we have learned enough to advise the planner to mimic the individual's choice.

Although the preceding assumption appears to finesse the identification problem, it rules out many plausible possibilities, e.g., that escaping responsibility enhances well-being. Also, there may be ambiguity and disagreement about the structure of the relevant meta-problem, S^I, with different candidates pointing toward different normative conclusions. To resolve that ambiguity objectively, we might consider defining S^I as the meta-problem that literally replaces the planner with the individual, but that approach has unacceptable implications. For example, in the context of time-inconsistent behavior, it would imply that policy prescriptions should depend on the timing of the *planner's* choice. If we think that time-inconsistency reflects psychological struggles with self-control (as many psychologists and economists do), it is peculiar to assume that the planner's timing affects the individual's well-being at all, let alone in the same way as the timing of the individual's choice (particularly given that the individual may not know when the planner chooses).

2.2 Allowing for divergences between preferences and behavior

Some proposals for behavior welfare analysis attempt to identify well-being by seeking rationalizations for behavior that weaken the relationship between choice and any single coherent preference relation, in most cases without broadening the preference domain. It is important to emphasize that, in confining attention to a narrow preference domain, one necessarily imposes a potentially objectionable identifying restriction. The desire to explain non-standard behavioral patterns is only part of the rationale for adopting a broadened preference domain. Intuition, introspection, and evidence from the field of psychology suggest that considerations such as temptation, regret, and anticipation actually do factor into experienced well-being. Clearly, if one allows for divergences between preferences and choices in addition to a broadened preference domain, the identification of well-being can only become more problematic. Consequently, the approach considered in this section does not escape the identification problems discussed in the previous section.

Existing papers within the pertinent literature employ two distinct strategies for relaxing the assumption that choices perfectly implement a single coherent preference relation. The first strategy is to offer a model of an imperfect decision process. With sufficient knowledge (or assumptions) concerning the manner in which that process maps preferences into choices, one can invert the mapping and recover preferences from choices. Examples include Bernheim and Rangel [2004],[5] Carmichael and MacLeod [2006], Koszegi and Rabin [2008b], Dalton and Ghosal [2008], and Manzini and Mariotti [2008]. The second strategy is to assume that choices reflect the interplay between two or more conflicting objectives. In that case, choices depend on the nature of the process that aggregates the objectives. The more one knows (or assumes) about that process, the more one can infer about the underlying objectives. Examples include Laibson et. al. [1998], O'Donoghue and Rabin [1999], Asheim [2008], Noor [2008a,b], and Green and Hojman [2008].[6]

The sheer number of papers cited in the previous paragraph reminds us that economists are notoriously adept at proposing many clever rationalizations for the same set of facts. In the current context, that cleverness is a curse, because different rationalizations for a given choice pattern typically have different normative implications. There is a danger that debates over conflicting rationalization-based welfare proposals will either devolve into potentially fruitless and irresolvable arguments over which representation is "correct," or simply cause the profession to throw up its collective hands in bewilderment. Thus, economists' efforts to achieve general acceptance of a rationalization-based normative framework may be doomed by our own boundless inventiveness.

Identification based on choice. Suppose we are interested in understanding choices within some large class of decision problems. Each potential problem requires the individual to choose from some set of objects; let $\mathbb{X}$ denote the universe of alternatives. We can represent

[5] We now believe it is more appropriate to interpret our work on addiction (Bernheim and Rangel [2004]) within the context of the welfare framework discussed in Section 4, which equates welfare with choice rather than with well-being, and I will say more about it in that context.

[6] The discussion below focuses on the identifiability of objectives. Even if it is possible to recover multiple conflicting objectives, one must still settle on a method of aggregating them for the purpose of conducting normative analysis. Some authors advocate the Pareto criterion (e.g., Laibson et. al. [1998]), while others favor a single revealed objective based on either intuition (e.g., O'Donoghue and Rabin [1999]) or philosophical arguments (e.g., Noor [2008a,b]). In my view, deference to one revealed objective rather than another reflects an arbitrary external subjective judgment.

nearly all of the decision-making models listed above (as well as many others) as a triplet, (I, γ, P), where $I = \{1, 2, ..., n\}$ is a set of objectives or motivations (possibly of different selves), P is a vector of preference orderings over $\mathbb{X}$ (one for each objective), and γ is a mapping that assigns, for each choice situation, a process that maps the objectives to outcomes.[7] Each behavioral theory, (I, P, γ), maps to a choice correspondence, C, defined on the set of decision problems; let Λ denote that mapping. If one imposes sufficient structure on I and γ, it may be possible to invert Λ and recover the preference vector from the choice correspondence. Consequently, some of the papers listed above claim to generalize the revealed preference paradigm. The central problem, however, is that preferences are only "revealed" if one accepts the process model, (I, γ), as a literal and accurate description of decision-making. Fixing I' and γ', for any given correspondence C', there may be only one preference profile P' for which $\Lambda(I', P', \gamma') = C'$, but there may be many theories of the form (I'', P'', γ'') that are observationally equivalent in terms of choice, in the sense that $\Lambda(I'', P'', \gamma'') = C'$, where $(I'', P'') \neq (I', P')$, and where (I'', γ'') and (I', γ') entail different intra-personal games, expectation formation processes, equilibrium notions, and/or conflict resolution rules. In that case, absent some compelling non-choice justification for accepting (I', γ'), preferences are either partially or completely unidentified.

I will use Γ to denote the set of "admissible" processes (including both the process mapping γ and the set of objectives I) that one might entertain to account for behavior. Define $\Pi(\Gamma, C) \equiv \{P \mid \Lambda(I, P, \gamma) = C \text{ for some } (I, \gamma) \in \Gamma\}$; if we restrict attention to Γ,

[7] For example, in Bernheim and Rangel [2004], I is a singleton set; P consists of a single preference relation; γ indicates that the individual will select either the maximal element of that relation or a fixed alternative (consumption of the addictive substance), depending on the choice situation. In the literature on quasihyperbolic discounting, I is usually defined to include one objective for each moment in time, P is the vector of preferences for those time-dated selves, and γ describes an intra-personal game, an assumption about expectations, and a notion of equilibrium for each choice situation. In Green and Hojman [2008], I indexes an arbitrarily large number of internal motivations, any two of which may or may not conflict, P describes each motivation, and γ is a scoring rule. (Formally, Green and Hojman focus on finite sets of potential alternatives, and model a population of motivations, λ, as a probability distribution over the set of possible strict orderings. As long as the elements of λ are rational, one can also model the population of motivations as a vector of orderings, as I suggest. If elements of λ are irrational, one can approximate the population of motivations arbitrarily well with a vector of orderings.)

then for any choice correspondence C, we can infer that $P \in \Pi(\Gamma, C)$.[8] We can state the conditional identification problem as follows: for a given choice correspondence C and set of admissible processes Γ, to what extent does $\Pi(\Gamma, C)$ narrow down possible preferences? Clearly, without *a priori* restrictions on Γ, the identification problem is insurmountable. Consider, for example, process models in which an ordering, P, describes well-being, but γ specifies that another entirely unrelated ordering, P', dictates choice. In that case, choices will reveal P', not P. Unless we treat such processes as inadmissible, preferences will be completely unidentified even if choice patterns satisfy standard axioms. Likewise, if we refuse to rule out masochistic decision processes that select the worst alternative rather than the best, it becomes possible to stand any theory on its head, and consequently impossible to distinguish good from bad. The preceding examples are, of course, contrived, and the reader may feel comfortable ruling them out. However, the critical questions still remain: on what basis does one draw the line between admissible and inadmissible models of decision processes (clearly, one cannot rely on choice patterns), where does one draw it, and what do the resulting restrictions imply about the identifiability of preference?

Illustration: the identification of multiple objectives. Within the pertinent literature, only Green and Hojman [2008] address the identification problem systematically while defining the set of admissible processes broadly. Their analysis illustrates and underscores how difficult it is to identify and evaluate well-being usefully once a reasonably wide range of non-standard explanations for choice is admitted. Specifically, they allow for the possibility that the individual has any number of objectives and uses any decision process that mimics a scoring rule (for example, plurality rule or the Borda rule). In one sense, their set of admissible processes, denoted Γ^{GH}, is quite inclusive, but it also entails meaningful restrictions. They do not offer evidence to justify those restrictions; rather, they interpret any particular model, (I, P, γ), as an as-if representation of choice. That interpretation would appear to undermine the normative force of their framework: if there are two theories, (I', P', γ') and (I'', P'', γ''),

[8] Note that $\Pi(\Gamma, C)$ may be empty.

with $(I', \gamma') \in \Gamma^{GH}$, $(I'', \gamma'') \notin \Gamma^{GH}$, and a choice correspondence C such that $\Lambda(I', P', \gamma') = \Lambda(I'', P'', \gamma'') = C$ and $P'' \notin \Pi(\Gamma^{GH}, C)$, it is difficult to understand the rationale for ruling out P'' upon observing C. After all, (I', P', γ') and (I'', P'', γ'') provide equally valid as-if representations of C.

Even limiting attention to Γ^{GH}, Green and Hojman's analysis implies that identification is problematic. To illustrate the difficulties, suppose for simplicity that there are only two potential alternatives, x and y. Each element of P must either rank x above y (ordering A) or y above x (ordering B); moreover, every scoring rule is equivalent to majority rule. Assuming the individual always chooses x over y, Green and Hojman would conclude only that a majority of elements of P (the cardinality of which is indeterminate) correspond to ordering A, and a minority to ordering B. That inference would not justify the conclusion that the individual is better off with x than y for at least two reasons: first, even if ordering B has less influence on decision making than ordering A, it may be more closely related to overall well-being; second, the rankings do not indicate the intensity of preference for one alternative over the other. Without cardinal information concerning preferences, Green and Hojman would therefore make only a weak statement concerning welfare: it is possible that x is unambiguously superior to y (in the sense that all objectives might agree), but not possible that y is unambiguously superior to x. In their view, standard ordinal welfare statements assert considerably more about well-being than one can logically infer from choice, even when standard choice axioms are satisfied, because they ignore the potential existence of "minority preferences" that do not express themselves through choice.

Illustration: time-inconsistency, part 1. Next I illustrate the identification problem in the context of a particular non-standard choice pattern: time-inconsistency. Consider the task of choosing a lifetime consumption vector, $c = (c_1, ..., c_T)$, where $c_t \geq 0$ denotes the level of consumption at time t; $\mathbb{X}$ consists of the set of all such vectors. A decision problem (X, τ) involves a set of lifetime consumption vectors, $X \subset \mathbb{X}$, and a decision tree, τ, for selecting an element of X. The decision tree describes the options available at each point in time (including precommitment opportunities), how those options depend on past actions, and how they affect the options that will be available in future periods. A choice correspondence maps decision problems to consumption vectors. The correspondence is time-inconsistent if $C(X,\tau)$ differs from $C(X,\tau')$ for some opportunity set X and two decision trees, τ and τ'.

One can account for many forms of time-inconsistency by assuming that decision making involves an intrapersonal game between time-dated selves, in which those selves have perfect foresight and play subgame-perfect equilibria. Formally, such models take the form (I^T, P, γ^T), where I^T indexes the time-dated selves, P is a vector of preference relations over lifetime consumption vectors (one relation for each self), and γ^T reformulates decision problems as intrapersonal games between time-dated selves. For the moment, I will restrict attention to models with the aforementioned features, thereby limiting the set of admissible processes to $\Gamma^T = \{(I^T, \gamma^T)\}$. As I explain below, the identification of preferences becomes even more problematic when one broadens the set of admissible processes beyond Γ^T.

Unfortunately, even if one assumes that (I^T, γ^T) correctly describes the decision process, preferences are only partially identifiable (because $\Lambda(\Gamma^T, \bullet)$ is only partially invertible): we can in principle recover the preferences of any time-dated self for current and future consumption, but choices cannot illuminate any aspect of preferences for past consumption.[9] Thus, if $\Pi(\Gamma^T, C)$ is non-empty, it contains many preference profiles. In the context of the well-known β,δ model of quasihyperbolic discounting (popularized by Laibson [1997] and O'Donoghue and Rabin [1999]), Laibson et. al. [1998] and others assume that each self is indifferent with respect to past consumption, but any other assumption would generate an observationally equivalent choice correspondence. Because the implications of many normative aggregation criteria (such as the Pareto criterion) are sensitive to which preference profile we select, that equivalence is potentially problematic.

Though always problematic, the identification issue described in the previous paragraph

[9] Formally, consider some vector of preferences P that corresponds to the utility functions $U_1(c), ..., U_T(c)$. Suppose there is a choice correspondence C for which $\Lambda(I^T, P, \gamma^T) = C$. Let $\phi_1(c), ..., \phi_T(c)$ be any collection of functions with the property that $\phi_t(c)$ is invariant with respect to c_k for $k \geq t$, and let P' be the vector of preferences corresponding to the utility funcitons $U_1(c) + \phi_1(c), ..., U_T(c) + \phi_T(c)$. Then $\Lambda(I^T, P', \gamma^T) = C$.

is not necessarily fatal. For example, following Bernheim and Rangel [2009], one can define a *robust multi-self Pareto optimum* as an alternative that is not Pareto improvable for any preference profile $P \in \Pi(\Gamma^T, C)$. That requirement is quite demanding: it acknowledges our ignorance concerning backward-looking preferences by requiring that no feasible option Pareto dominates the alternative in question regardless of how any self feels about past consumption. In many settings, the set of robust multi-self Pareto optima may prove to be empty. However, for the β,δ model, Bernheim and Rangel show that a consumption trajectory is a weak multi-self Pareto optimum within a standard intertemporal budget set if and only if it maximizes decision utility at the first moment in time. If that first moment is short, weak multi-self Pareto optima approximately coincide with the choices that are optimal according to the long-run criterion.

Falsifiability versus identifiability. Some of the papers within this literature focus on the question of whether a process model is falsifiable based on choice patterns (e.g., Manzini and Mariotti [2008], Dalton and Ghosal [2008]). They urge us to formulate a clearly testable model of economic choice, derive its welfare implications, and – as long as the model is not rejected by the data – use those implications to guide normative analysis. The problem with that agenda is that it does not come to grips with the issue of observational equivalence. Formally, for any class of admissible processes Γ, let $\Omega(\Gamma)$ denote the set of associated choice correspondences; that is, $C \in \Omega(\Gamma)$ iff there is some process $(I, \gamma) \in \Gamma$ and preference vector P such that $\Lambda(I, P, \gamma) = C$.[10] The theory $(I, \gamma) \in \Gamma$ is falsifiable if there is a conceivable choice correspondence $C \notin \Omega(\Gamma)$. As the accumulating evidence on an individual's choices points increasingly toward some $C^* \in \Omega(\Gamma)$, a proponent of the theory "$(I, \gamma) \in \Gamma$" can claim that the absence of falsification validates the use of $P \in \Pi(\Gamma, C^*)$ for normative analysis. And yet, a proponent of any other theory "$(I, \gamma) \in \Gamma'$" for which $C^* \in \Omega(\Gamma')$ (where Γ' is some other class of processes) can make the same claim concerning $P' \in \Pi(\Gamma', C^*)$. That state of affairs is plainly problematic when $P \neq P'$. To appreciate the severity of the

[10] Plainly, each element of the preference vector must be an ordering over $\mathbb{X}$.

problem, consider the fact that *every conceivable* choice correspondence has many multiple-objective representations (Green and Hojman [2008]). Certainly, the absence of choice-based falsification for any particular multiple-objective representation of C^* is not a valid justification for using that representation, rather than some other representation of C^*, as the basis for normative judgments.

If one is willing to maintain the hypothesis that all human beings employ the same basic type of decision process (but potentially differ with respect to preferences), then it becomes possible to differentiate between two classes of processes, Γ and Γ', provided that $\Omega(\Gamma') \neq \Omega(\Gamma)$. Specifically, if $\Omega(\Gamma)$ contains the choice correspondences for every member of a population while $\Omega(\Gamma')$ does not, one might be inclined to reject Γ' in favor of Γ, even for individuals whose correspondences lie in $\Omega(\Gamma') \cap \Omega(\Gamma)$.

There are, however, two problems with that version of the falsifiability agenda. First, the maintained hypothesis may be incorrect. Indeed, there is evidence that different people exhibit activation in different regions of brain when making the same decisions, and those differences are related to their choices. Individuals with choice correspondences in $\Omega(\Gamma') \cap \Omega(\Gamma)$ may well use the type of process found in Γ', while those with correspondences in $\Omega(\Gamma) \backslash \Omega(\Gamma')$ use the type of process found in some other set, Γ''. In that case, only simplicity arguably favors Γ over $\Gamma' \cup \Gamma''$.

Second, for any process theory, Γ, it is usually possible to formulate other theories simply by reinterpreting and/or relabeling elements of Γ (most notably, the elements that are labeled as preferences). For any such alternative theory, Γ', we necessarily have $\Omega(\Gamma') = \Omega(\Gamma)$, so that the two theories are observationally equivalent based on choice. And yet, for any given $C \in \Omega(\Gamma') = \Omega(\Gamma)$, we may discover that $\Pi(\Gamma, C)$ and $\Pi(\Gamma', C)$ are disjoint. The β, δ model provides an excellent illustration.

Illustration: time-inconsistency, part 2. Consider any choice correspondence $C^{\beta\delta}$ that conforms to quasi-hyperbolic discounting for a particular β, δ, and flow utility function u. Staying within Γ^T, we know that $\Pi(\Gamma^T, C^{\beta\delta})$ includes any vector of preferences P

corresponding to utility functions of the form $\phi(c_1, ..., c_{k-1}) + u_k(c_k) + \beta \sum_{t=2}^{T} \delta^{t-k} u(c_t)$ $(k = 1, ..., T)$ (see above). Outside of Γ^T, there are many other possibilities. All of the following are observationally equivalent process models, and each has different normative implications. (1) There is a unitary self ($I = \{1\}$), and P corresponds to a single time-consistent utility function, $\sum_{t=1}^{T} \delta^{t-1} u(c_t)$, but irrational attraction to immediate prospects ("present bias") in each period k causes the decision maker to mistakenly inflate the importance of current flow utility, and to maximize $u_k(c_k) + \beta \sum_{t=k+1}^{T} \delta^{t-k} u(c_t)$.[11] (2) There is a unitary self, and P corresponds to a single time-consistent utility function, $\sum_{t=1}^{T} (\beta\delta)^{t-1} u(c_t)$, but irrational anxiety over distant prospects ("future bias") in each period k causes the decision maker to mistakenly inflate the importance of flow utility enjoyed after period $k+1$, and to maximize $u_k(c_k) + \beta \sum_{t=k+1}^{T} \delta^{t-k} u(c_t)$. (3) There are two selves ($I = \{1, 2\}$), a "planner" and a "doer." P assigns a time-consistent utility function, $\sum_{t=1}^{T} \delta^{t-1} u(c_t)$, to the planner, and a perfectly myopic objective function ($u(c_t)$ in period t) to the doer. The planner and the doer bargain over choices and renegotiate at each moment in time; β is the planner's bargaining weight.[12] (4) There are multiple selves, one for each node in any given decision tree.[13] Each self has β, δ preferences defined over the levels of current and future consumption that might emerge if its node were reached. As Asheim [2008] demonstrates, when we associate selves with nodes rather than dates, application of the Pareto criterion leads to strikingly different conclusions concerning welfare. (5) If one is willing to redefine the objects of choice for the set of decision problems under consideration, there is an observationally equivalent process model involving a unitary self with "temptation preferences" (Gul and Pesendorfer [2001]) defined over both consumption vectors and the opportunity sets from which they are chosen; see Krusell, Kuruscu, and Smith [2001] and Dekel and Lipman [2007] for details.

[11] With this reinterpretation of the β, δ model, δ remains a preference parameter, while β becomes a feature of the decision process.

[12] At any point in time k, intraself regengotiation will lead the decision maker to maximize $\beta \sum_{t=k}^{T} \delta^{t-k} u(c_t) + (1-\beta) u(c_k) = u(c_k) + \beta \sum_{t=k+1}^{T} \delta^{t-k} u(c_t)$. Notice that one can also think of the planner and the doer as spouses. Thus, marriages between time-consistent individuals with differing degrees of impatience can lead to time-inconsistent decisions (see Bernheim [1999]).

[13] This possibility falls outside of the general framework described above, because the definitions of I and P depend upon the decision problem.

Here, there falsifiability agenda is plainly ill-conceived, because choice-based tests have zero power against many pertinent alternatives. As long as the β, δ model survives empirical tests involving choice data, proponents of the standard multi-self interpretation (e.g. Laibson et. al. [1998]) could point to the absence of falsification as validating the associated normative criterion, but so could a proponent of any alternative interpretation.

Identification involving non-choice evidence. Some of the papers in this literature acknowledge that any justification for a restriction on Γ must be found in non-choice evidence. However, given the limited state of knowledge concerning decision processes, it is extremely challenging to justify a general process model as literally correct within a broad domain.[14] At this juncture, general models of decision making that involve specific choice procedures and/or intrapersonal games are most appropriately regarded as metaphors rather than literal descriptions. While those metaphors are useful from the perspective of positive analysis, normative analysis requires us to take them more seriously and literally than the available evidence merits. Impressionistic evidence from psychology and neuroscience may motivate certain models of process, but in my view it falls well short of justifying the types of restrictions that one must adopt to validate the welfare criteria advocated, for example, by Dalton and Ghosal [2008] and Manzini and Mariotti [2008].[15]

From a normative perspective, the most vexing questions about decision processes are *interpretive*, not mechanical. According to many behavioral models, decisions reflect the interplay between well-behaved preferences and other motivational systems. For example, Manzini and Mariotti [2008] examine a two-step model of choice in which the individual first eliminates objects belonging to "psychologically shaded" categories, and then chooses from among the remaining objects based on a "preference" relation. In Cherepanov et. al. [2008], the individual first restricts attention to objects that are best according to at least one of

[14] See Bernheim [2009] for a general discussion concerning the potentialities and limitations of neuroeconomic evidence.

[15] Thus, Dalton and Ghosal's claim that certain behavior patterns reflect suboptimal choices as judged by the decision maker's true objectives is an interpretation, not a well-supported inference.

many "rationales," and then selects the "most preferred" object from that set.[16] Yet one could relabel the psychological shading relation or the rationales as "preferences" and reach different normative judgments. Psychological evidence can in principle identify motivations, and neural evidence can isolate motivational systems, but I do not see how either can help economists decide which are preferences and which are not.

The problem is readily apparent in the context of the β, δ model. I have described six distinct interpretations of that model, three with unitary selves, one with a dual self, and two with multiple selves, each of which has different normative implications. Because the distinctions between those interpretations largely involve labeling rather than the mechanics of choice, it is not at all apparent that non-choice evidence can discriminate between them. Consider, for example, the finding that the prospects of immediate and delayed payoffs lead to activation in different regions of the brain; following McClure et. al. [2004], I will refer to those as the "β" and "δ" regions. In a unitary-self setting, activation in the β region might reflect "present bias" or "temptation," or activation in the δ region might reflect "future bias." In a dual-self setting, the β and δ regions might correspond to the doer and the planner. In a multiple-self setting, β and δ activity might simply reflect the process by which a self assesses a β, δ valuation at a particular point in time or node in a decision tree. As for resolving the dispute between Laibson et. al. [1998] and Asheim [2008] as to whether selves are associated with dates or nodes, I am at a loss to see how non-choice evidence could be helpful.

Despite the preceding comments, I do see a useful role in normative analysis for non-choice evidence, at least in limited circumstances (such as the contexts discussed by Bernheim and Rangel [2004] or Koszegi and Rabin [2008b]). I will explain and justify those uses in Section 4.9.

[16] As Manzini and Mariotti [2008] observe, based on choice evidence alone, the "rationalization" model of Cherepanov et. al. [2008] is observationally equivalent to a simple version of Manzini and Mariotti's [2008] "categorize-then-choose" theory; both are characterized by the same choice axiom (WWARP).

3 Welfare as measured well-being

Another proposal for behavioral welfare analysis is to measure well-being directly, as urged by Kahneman et. al. [1997], Kahneman [1999], Frey and Stutzer [2002], Kahneman and Sugden [2005], Layard [2005], and Koszegi and Rabin [2008a], among others. Welfare analysis might build upon a sizeable body of work in psychology concerning the measurement of happiness, satisfaction, and/or other related concepts (all of which I will henceforth subsume under the heading of "happiness" for the sake of brevity). Happiness research has already achieved a toe-hold in economics; examples include Gruber and Mullainathan [2002] , Frey and Stutzer [2004], Kimball and Willis [2006], and Kimball et. al [2006]. Yet it is one thing to assert that economists ought to study happiness, and quite another to suggest that measures of happiness can provide a rigorous foundation for welfare analysis, either alone or in concert with indicia of neurobiological activity. While I strongly endorse the first statement, I have profound doubts concerning the second. In this section, I explain the conceptual basis for those doubts.

Well-being versus self-reported happiness. Happiness is commonly understood as an internal state of mind. However, psychologists (and some economists) often define it operationally as the answer to a question such as "*On a scale from one to seven, where one is extremely unhappy and seven is extremely happy, how do you feel right now?*" To avoid confusion, I will henceforth refer to the internal state of mind as "well-being," and to the operational concept as "self-reported happiness." Unfortunately, much of the relevant literature treats these distinct concepts as if they were equivalent. By failing to draw a clear distinction between them, even the most thoughtful commentators succumb to misleading language. For example, Kimball and Willis [2006] write: "...some economists think happiness can't be measured well. *This is just not true.* Happiness (current affect) is one of the easiest of all subjective concepts to measure." On the contrary, only *reports* of happiness are easy to measure. No one has yet proposed any sensible way to measure an individual's internal state of well-being directly. Granted, we can try to infer that state from self-reported happiness and/or neurobiological activity, but the validity and accuracy of any inferential procedure must be evaluated according to the same standards used elsewhere in economics. As I explain below, it is every bit as problematic to identify useful information concerning internal well-being from data on self-reported happiness and/or neurobiological activity as it is from choice.

In Section 4, I will argue that a choice-based approach to welfare analysis can finesse the identification problem by equating welfare with choice rather than well-being. Likewise, one could simply equate welfare with self-reported happiness. However, one must then be willing to accept many highly problematic implications. Suppose you are inclined to report your happiness as average. If someone points a gun at your head and states (credibly) that they will shoot you unless you say you are extremely happy, no doubt you will comply, even though the physical threat makes you considerably less happy. Creating the threat causes your internal well-being and self-reported happiness to move in opposite directions.[17] Yet if we equate welfare with self-reported happiness, we must conclude that the threat improves welfare. This example, while extreme, has practical counterparts. Citizens of countries with oppressive governments are often reluctant to express dissatisfaction or unhappiness with public policies lest they face persecution. Manifestly unpopular despots have been known to defend their regimes by citing evidence of overwhelming support from sham elections. Yet we would presumably disagree with the claim that those regimes are welfare-improving.

At first, this problem might seem to have a straightforward solution: equate welfare with happiness reported under non-coercive conditions. Unfortunately, the issue is not so simple. The implicit justification for this solution is that coercion induces people to misreport their internal states of well-being. That justification is pertinent only if we abandon our original position and equate well-being with the internal state, rather than with the report. In addition, coercion is simply a extreme form of influence. Similar conceptual problems arise

[17] A choice-based approach avoids that difficulty. Suppose someone chooses A over B when left alone, and B over A when threatened. In that case, we would conclude that he prefers A to B, but also prefers "B and life" to "A and death." Notably, we do not conclude that he is better off with "B and life" when threatened than with A when not threatened.

in less dramatic contexts. Suppose a policy reduces personal income by five percent while fostering a social norm that encourages people not to complain; surely, it would be incorrect to conclude that they are necessarily better off. Thus, to implement the solution proposed at the outset of this paragraph, one would require a clear criterion for judging whether "excessive" influence is present. It is difficult to imagine a criterion that would neither implicitly nor explicitly reference consequences that induce the misreporting of true well-being. In short, unless we accept reported happiness unconditionally, we are inevitably compelled to define welfare in terms of an internal state.

The aggregation problem. It is natural to conceptualize the internal state of happiness as a vector encompassing the many aspects of mental and physical well-being. For the purpose of normative analysis, we would like to find a measure of happiness that appropriately aggregates that multi-dimensional state into a scalar. If the human brain routinely performs that aggregation, we might hope to measure an aspect of the internal state that summarizes overall happiness in single dimension. Yet there is no guarantee that such aggregation actually takes place. (Even if it does, we would be hard-pressed to distinguish the aspect that summarizes overall well-being from those that contribute to well-being.) In that case, to construct a useful welfare measure based on internal states, economists would need to settle on a method of aggregation. Justifying any such method rigorously and objectively is likely to prove problematic.

The use of self-reported happiness poses similar problems. When we are asked about our happiness, we may simply examine our many sensations and aggregate them in an unsystematic and arbitrary way. Furthermore, our answers are typically limited in scope. One can express happiness with an event, with life in general at a moment in time, or broadly with one's current situation and future prospects. To be useful, any question concerning happiness must explicitly identify the intended scope.[18] If the scope of the question is narrower than the individual's entire life, one cannot evaluate welfare without aggregating

[18] In fact, the scope of the typical question is murky, and unambiguous questions are difficult to formulate. However, those are practical issues, and my focus here is conceptual.

responses. For example, if we assess momentary happiness (*how do you feel right now?*), we must aggregate responses over time. In the absence of objective principles, the choice of any particular aggregator is inherently arbitrary.

It is far from clear that we escape this problem by asking apparently all-encompassing questions (e.g., concerning "life satisfaction"). I see no way to test the hypothesis that an answer omits nothing of consequence; that proposition must simply be an article of faith. Moreover, an individual may respond differently to such questions at different points in her life. To some extent, those discrepancies may reflect differences in information, but they may also reflect differences in perspective. If someone is twenty-five, should we evaluate her well-being based on a twenty-five year-old's perspective on life, or a fifty year-old's perspective? Either we must justify the use of one particular perspective, or we must aggregate across perspectives and justify the aggregator. Because any given question elicits an answer at a single point in time, it cannot aggregate perspectives for us.

The identification problem. Let's define an *environment*, E, to consist of the external processes that govern experiences, as well as all welfare-relevant preconditions. An environment may or may not present an individual with future decisions, and the preconditions may include past choices. An environment maps to a vector of sensations, $s = S(E)$, which may describe either physical states (such as hunger and fatigue), abstract emotions (such as shame or relief), or feelings induced by expectations of future sensations. Internal well-being, u, aggregates those sensations: $u = V(s, E)$. I include E as an argument of U because the perception of an environment is itself a sensation, one that may color hedonic experience. I allowed for similar possibilities in Section 2.1 by permitting well-being to depend on the choice set (an aspect of the environment). Whether or not V depends directly on E, we can write well being as a function of the environment and nothing else:

$$u = U(E) \equiv V(S(E), E). \tag{2}$$

The mapping U is not observed. Our central objective is to identify it up to a monotonic transformation, so we can rank environments according to the well-being they generate.

Let h denote self-reported happiness. As discussed in detail below, the report may depend on well-being, sensations, and/or the environment: $h = R(u, s, E)$.[19] Notice that we can write h as a function of E alone:

$$h = H(E) \equiv R(U(E), S(E), E). \tag{3}$$

By assessing self-reported happiness in a wide range of environments, we can observe the mapping H. We may also observe components of the mapping S. The question before us is whether we can recover the mapping U from that information.

Having formulated the problem in this way, it should be obvious that identification is hopeless without further restrictions. In particular, for any mappings U, H, and S, there exists functions R and V such that equations (2) and (3) hold. For example, R and V might depend only on E, in which case one can simply take $V(E) = U(E)$ and $R(E) = H(E)$. Even if we insist that well-being is responsive to sensations and that self-reported happiness is responsive to well-being, we are no better off.[20]

Are there stronger but nevertheless reasonable assumptions that would permit us to recover U? I will attack that question in two steps. First, I will ask whether there is any hope of recovering U from H alone. I will then investigate the possible benefits of incorporating information concerning S.

Identification of well-being from self-reported happiness. Those who advocate the use of self-reported happiness as a welfare measure implicitly assume that greater well-being leads to higher levels of reported happiness (monotonicity of R in u), and that those reports do not depend directly on either sensations or the environment (invariance of R with respect to s and E). In that case, rankings according to u and h plainly coincide, and H identifies U up to a monotonic transformation (R). The issue is whether those identifying assumptions

[19] In principle, u could also depend on h. Indeed, if answers to questions about happiness do not affect well-being, people would have no reason to report happiness accurately. That issue does not arise if the individual is asked about his recent happiness, and if the question is unexpected.

[20] To see why, consider the possibility that V and R might be separable: $V(s, E) = V_1(s) + V_2(E)$, and $R(u, s, E) = R_1(u)+ R_2(s) + R_3(E)$. For any U, H, S, V_1, R_1, and R_2, one can simply take $V_2(E) = U(E) - V_1(S(E))$ and $R_3(E) = H(E) - R_1(U(E)) - R_2(U(S(E)))$.

are reasonable. In my view, they are not.

The assumption that R is invariant with respect to s rules out the reasonable possibility that, when reporting happiness, an individual may focus disproportionately on special aspects of well-being, and potentially ignore others. Most people may construe "happiness" as encompassing limited aspects of well-being. For example, they may neglect to consider the satisfaction derived from speaking truthfully, behaving honorably, or adhering to some moral code. One can attempt to overcome that difficulty by asking broad questions about life satisfaction, but ultimately I see no way to confirm that the scope of the question has been defined and interpreted appropriately.

The assumption that R is invariant with respect to E is also untenable, for two reasons. First, as I have already discussed, aspects of the environment (especially coercion and influence) plainly affect an individual's willingness to report happiness truthfully. Second, there is no absolute scale for measuring happiness. When we ask people to express their happiness using a unitless scale (e.g., one to seven), we compel them to invent their own normalizations; they decide for themselves what each potential response signifies. The chosen normalization may well depend on aspects of the environment. For example, people may use the mid-point of the scale to denote their typical state of happiness. Psychologists attempt to address such concerns by employing "debiasing" procedures, but there is no absolute standard against which one can evaluate bias; hence, the effectiveness of such procedures is inherently subjective and unprovable.

Allowing for even the most simple and natural relationships between R and E seriously undermines our ability to draw meaningfully inferences about well-being from self-reported happiness. Consider the Easterlin Paradox, which consists of the observation that increases in income and consumption over time do not lead to greater self-reported happiness (Easterlin [1974, 1995, 2003]). The paradox is sometimes construed as evidence that economic growth does not improve well-being, but that conclusion is unwarranted. Suppose $R(u, s, E) = m + u - \overline{u}$, where m is the midpoint of the happiness scale and $\overline{u}$ is the expected level of well-being in the environment E; in other words, people normalize their reports so that the scale's midpoint corresponds to average well-being. In that case, Easterlin's finding is consistent with any conceivable relationship between income and average happiness.

Next consider the phenomenon of "projection bias." Following a dramatic and permanent change in circumstances (e.g., winning the lottery, failing to obtain tenure), reported happiness often changes initially, and then converges back toward its original level over time. Yet most people predict that such events will have substantial permanent effects on their happiness (see., e.g., Loewenstein et. al. [2003]). The pertinent literature interprets those findings as reflecting a failure to anticipate hedonic adaptation. Yet if one assumes that people (a) renormalize the happiness reporting scale based on recent experience, and (b) report expected future happiness based on the scale currently in use, rather than the scale they will use once the future arrives, one can explain the same set of facts while positing zero hedonic adaptation and rational expectations.

Finally, consider any event that produces a favorable shift in reported happiness, in the sense of first-order stochastic dominance (FOSD). Without highly restrictive assumptions about the function R, one cannot rule out the possibility that the event unambiguously reduces well being. To illustrate, let's suppose that R takes the form $R(u, s, E) = Q(u) - \alpha Q(\overline{u})$ where Q is an increasing function, and $\overline{u}$ is once again the expected level of well-being in the environment E. In that case, even assuming $\alpha < 1$, one can construct reasonable examples in which a downward FOSD shift in the distribution of well-being leads to an upward FOSD shift in self-reported happiness.[21] Thus, for example, Gruber and Mullainathan's [2005] finding that higher cigarette tax rates lead to greater reported happiness among smokers proves nothing about smokers' well-being.

Identification of well-being from neurobiological activity. So far, I have assumed that sensations, s, are unobservable. With the emergence of new noninvasive techniques for monitoring neurobiological activity, such as fMRI scans, we can detect the neural manifesta-

[21] To construct such an example, assume that Q is S-shaped and that u has binary distribution. Because happiness is elicited on a scale with an upper and lower bound, S-shaped reporting functions are plausible.

tions of those sensations. Few people would dispute the claim that such activity is related to well-being. However, it is not at all obvious that the ability to observe neural activity overcomes the identification problems that preclude us from measuring well-being reliably.

Measures of neurobiological activity are of little use without additional information. Regions of the brain are not etched with functional labels. Without external correlates (such as expressions of happiness or choices), it would be impossible to say whether heightened activity in a particular portion of the brain reflects pleasure, pain, or something else entirely. Formally, knowledge of the mapping S cannot by itself inform our understanding of the mapping, V, that aggregates the components of s. Without further information, we cannot determine whether a given sensation is positive, negative, or neutral; nor can we assess the relative importance of different sensations in determining well-being.

A purely neurobiological measure of well-being would also raise vexing philosophical concerns. Suppose a scientist invents a pill that replicates the full spectrum of sensations associated with a two week vacation touring exotic locales, and leaves the person who swallowed it with the conviction that the perceived experience was real. Should we accept the implication that the pill and the vacation create the same level of well-being? I suspect that many if not most people would, if given a choice, select the vacation over the pill, on the grounds that real experience inherently trumps a simulation. I see no valid rationale for overruling that stated and (hypothetically) revealed preference. Instead, I am led to the conclusion that well-being likely encompasses more than identifiable sensations.

The key question, then, is whether one can use neurobiological measurements *in concert* with self-reported happiness to identify well-being. Contributors to the happiness literature have informally suggested that possibility. Specifically, they argue that correlations between self-reported feelings, biometric variables, and neural measurements corroborate the use of such objects as indicia of well-being (see, e.g., Larsen and Fredrickson [1999]). But that argument is circular; it demonstrates only that the variables in question have something in common, not that they individually or collectively embody true well-being. Formally, we have already seen that, if one permits the mappings V and R to depend directly on E and S, then true well-being is unidentified. That observation subsumes cases in which $H(E)$ and elements of $S(E)$ are highly correlated.

Still, one might hope that knowledge of the mappings H and S would permit the identification of U under less restrictive assumptions than knowledge of H alone. We have already seen that one can identify U from H under the restriction that R is invariant with respect to s and E (assuming plausibly that R is monotonic in u). Consequently, knowledge of $S(E)$ would be useful only if it allows us to relax those invariance restrictions. Even knowing $S(E)$, we clearly cannot identify $U(E)$ unless we impose strong *a priori* restrictions on the manner in which $R(u, s, E)$ varies with both s and E. Intuitively, if correlations between s and h are potentially attributable either to the direct effect of s on h or the effect of E on h rather than the effect of s on u, they reveal nothing about the relationship between u and h.[22]

What types of restrictions might suffice? If we assume that V is invariant with respect to E, information concerning s and h could shed light on the dynamics of renormalization. Suppose, for example, that a permanent change in the environment leads to a permanent change in s, but only a temporary change in h. One could then conclude that mean reversion in h reflects renormalization, not hedonic adaptation. On the other hand, if the change in s is also temporary, one would interpret the evidence as pointing toward hedonic adaptation rather than renormalization. If we assume in addition that R is invariant with respect to s and depends on E only through the scaling norm used to report happiness (as above), then the function $V(s)$ (and hence $U(E)$) is potentially recoverable up to a monotonic transformation.

Unfortunately, all of those restrictions are objectionable. Because I have already discussed the restrictions on R, I'll focus here on the assumed invariance of V with respect to E, which allows us to distinguish renormalization from hedonic adaptation. As circumstances

[22] That intuitive argument is easily formalized.

change, the brain may adapt its stimulus-response sensitivity to achieve maximal discernment among outcomes over the range of likely experiences (as in the evolutionary theory of Rayo and Becker [2007]). However, other portions of the brain may “interpret” the intensity of neural responses in light of the anticipated experiential range, so that the level of well-being for any given sensation s depends on expectations for the pertinent environment, E. In that case, mean-reversion in s following a permanent change in the environment could reflect a process of *internal* renormalization, rather than an actual change in experienced well-being. Despite their apparently divergent implications for welfare, those two hypotheses may not be empirical distinguishable.

In practice, our knowledge of the mapping S will be imperfect, rendering reliable identification of well-being even more difficult. At least two additional problems are likely to arise. First, the neural circuitry that registers sensations may also be involved in other cognitive functions, in which case measured sensations may contain spurious components. One can of course assume that those components amount to random noise, but it is also easy to imagine that they are systematically related to features of the environment, such as complexity. Second, there is no guarantee that any given list of sensations will encompass all welfare-relevant activity. Consequently, there is a risk that sensation-based welfare analysis would overlook key aspects of well-being, and potentially focus disproportionately on sensations that are most easily measurable. Economists might unintentionally overemphasize physical sensations such as hunger and fatigue while underemphasizing abstract sensations such as pride and shame. Such omissions can potentially confound the identification strategies discussed above. For example, if measured sensations exhibit systematically different degrees of regression to the mean than unmeasured sensations, estimates of renormalization dynamics will be unreliable, even under the restrictive assumptions listed above.

4 Welfare as Choice

As I mentioned in the introduction, an alternative interpretation of standard welfare economics holds that welfare is *defined* in terms of choice rather than underlying objectives. That perspective has a long intellectual tradition; see Little [1949].[23] Naturally, one must explain why *choice*, per se, is normatively compelling even though its relationship to well-being may be somewhat obscure and potentially imperfect. Two types of justifications are available. As Little observed, "[o]ne could... say that a person is, on the whole, likely to be happier the more he can have what he would choose. Or, alternatively, one can say that it is a good thing that he should be able to have what he would choose." I will refer to the first type of justification as *instrumental*, and the second as *non-instrumental.*

Instrumental justifications provide reasons to expect that a policy of respect for choice will, for the most part, promote well-being. One version holds that while choice may be an imperfect proxy for well-being, no better proxy is available. While I see merit in that argument, I acknowledge that advocates of happiness measures may disagree. Moreover, because an individual's internal state of well-being is not directly measurable, the issue may be impossible to settle. A second related instrumental justification holds that *unambiguous* choice is a reliable proxy for well-being. In other words, if someone consistently chooses option A over option B, and we cannot induce him to choose B over A without misleading, confusing, or coercing him, then we can safely conclude that he is better off with A than with

[23] In his seminal paper on revealed preference, Samuelson [1939] wrote that his object was to "develop the theory of consumer's behavior freed from any vestigial traces of the utility concept," but he did not explicitly address the topic of welfare. Sen [1973] disavowed the choice-centric perspective, writing that the"rationale of the revealed preference approach lies in the assumption of revelation and not in doing away with the notion of underlying preferences, despite occasional noises to the contrary." (Notably, he also criticized the revelation assumption.) He reasoned that the consistency axioms upon which economists usually rely have no force unless one assumes that preferences actually drive choices. Sen's argument ignores the fact that certain classes of mechanistic decision processes can generate choice correspondences that satisfy the same consistency axioms (see, e.g., Mandler, Manzini, and Mariotti [2008]). His argument is particularly weak in the context of behavioral economics. The assumptions underlying many behavioral models are chosen because they match empirical choice patterns, not because they reflect some potentially plausible if stylized conception of the decision process. Moreover, as explained in Section 2.2, when explicit models of cognition are proposed, the concept of preference is usually ambiguous.

B. Naturally, one can contrive counterexamples, and the inference could be wrong in a given context, but unambiguous choice may nevertheless create a strong presumption concerning well-being. I acknowledge that the presumption is largely based on introspection; once again, direct validation of the proxy is problematic. A third instrumental justification – one that invokes the tradition of liberalism – holds that a society committed to the principle of respect for choice will promote well-being more effectively than societies that permit those entrusted with governance to ignore or second-guess choices. That view was famously articulated by John Stuart Mill [1869], who wrote that an individual "... cannot rightfully be compelled to do or forbear because it will be better for him to do so, because it will make him happier, because, in the opinion of others, to do so would be wise, or even right." One can object that the argument presupposes an ability to measure well-being so that its level can be compared across societies, but for such judgments precise measurement is not necessarily required (and indeed, traditional utilitarian arguments for liberalism do not rely upon it).

While instrumental justifications portray choice as deriving its normative significance from its connection to well-being, non-instrumental justifications maintain that choices are normatively compelling simply because they are choices; hence it is possible to define welfare in terms of choice without implicitly invoking other objectives. That view finds support in the writings of certain libertarian philosophers, such as Robert Nozick [1974], who regard self-determination as a fundamental right rather than a means to an end. Libertarians typically construe that principle as constraining the legitimate scope of government. Nozick himself saw taxation as morally illegitimate, and argued for a minimal state charged only with protecting individuals from coercion. To the extent one adopts the comparatively moderate position that more extensive government activity is either desirable or inevitable, the principle of self-determination arguably implies that those involved in governance should judge the impact of interventions on individuals according to the choices those individuals would have made for themselves.

My work with Antonio Rangel (Bernheim and Rangel [2007, 2008, 2009]) adopts this choice-centric interpretation of standard welfare analysis, and develops a generalized welfare criterion that respects choice directly, without reference to the decision maker's underlying objectives.[24] We thereby finesse the thorny problems associated with formulating and justifying rationalizations for behavior, and avoid the intractable identification problems discussed in Section 2. Naturally, operationalizing the principle of respect for choice is problematic when choices conflict. However, useful behavioral theories do not imply that choice conflicts are ubiquitous. On the contrary, those theories are generally motivated by the observation that choice patterns exhibit a substantial degree of underlying coherence. We take that observation as our central premise, and devise welfare criteria that respect the coherent aspects of choice. Specifically, we propose replacing the standard revealed preference relation with an *unambiguous choice* relation: roughly, x is (strictly) unambiguously chosen over y (written xP^*y) iff y is never chosen when x is available. That criterion instructs us to respect choice whenever it provides clear normative guidance, and to live with whatever ambiguity remains. Though P^* need not be transitive, it is always acyclic, and therefore suitable for rigorous welfare analysis. The justification for this particular welfare relation is straightforward: any weaker criterion would fail to respect an unambiguous choice pattern, and (as explained below) any stronger criterion would necessarily overrule choice in some circumstances by deeming an object improvable within a set from which it is chosen.

In this section, I discuss the attractive features of our framework. Like standard welfare economics, it requires only information concerning the mapping from environments to choices; hence, its implementation does not require new types of data or breakthroughs in our understanding of decision processes. Because it encompasses any theory that generates a choice correspondence, it is applicable irrespective of the processes generating behavior, and regardless of whether one adopts a positive model that is preference-based, algorithmic, mechanistic, or heuristic. Indeed, it permits us to conduct welfare analysis even if "true

[24] Chambers and Hayashi [2008] also adopt this perspective. However, their approach, which involves the derivation of restrictions on welfare orderings from proposed axioms concerning the mapping from choice data to orderings, bears little resemblance to ours in other respects.

preferences" do not actually exist (as in Mandler, Manzini, and Mariotti [2008]). Moreover, because our welfare prescriptions are entirely choice-driven, they are unaffected by the particular rationalization one offers for a given choice pattern. Our framework generalizes standard choice-based welfare economics in two senses; first, the approaches are equivalent when standard choice axioms hold; second, for settings in which departures from those axioms are minor, our approach implies that one can approximate the appropriate welfare criterion by ignoring choice anomalies entirely. Our framework generates natural counterparts for the standard tools of applied welfare analysis and permits a broad generalization of the first welfare theorem. It is easily applied in the context of specific positive theories, and in some cases leads to novel normative implications. It has also been successfully applied in the context of tax illusion (Chetty, Looney, and Croft [2008]).

Naturally, when choice patterns are highly ambiguous, our welfare criterion is not especially discerning. However, our framework also provides a logical avenue for refining the criterion by placing restrictions on the welfare-relevant choice domain. In principle, one can justify such restrictions by marshalling non-choice evidence concerning choice processes. The most promising possibility is to seek evidence that particular choices involve errors in processing factual information; we propose classifying those choices as "mistakes" and excluding them from consideration. One can in principle justify restrictions based on other criteria as well.[25] We see choice-domain restrictions as a general strategy for incorporating welfare-relevant non-choice information, one that can potentially subsume many other approaches to behavioral welfare economics. For example, if (contrary to my expectations) adequate evidence is assembled to defend a normatively substantive process model, within our framework the same evidence would typically justify a choice-domain restriction that leads to the same welfare criterion.[26]

In this section, I briefly describe our framework and elaborate on some of its attractive

[25] For example, one could appeal to ethical principles as in Noor [2008a,b].

[26] As I explain below, in any given instance, some reinterpretation of the choice correspondence might be required.

features. I refer the interested reader to Bernheim and Rangel [2009] for additional details.

4.1 A general framework for describing choices

Let $\mathbb{X}$ denote the set of all possible choice objects. The elements of $\mathbb{X}$ need not be simple consumption bundles; for example, they could be lotteries, intertemporal outcome trajectories, or even consumption trajectories that depend on random and potentially welfare-relevant events. We define a *generalized choice situation* (GCS), $G = (X, d)$, as a constraint set, $X \subset \mathbb{X}$, paired with an *ancillary condition*, d.[27] An ancillary condition is an exogenous feature of the choice environment that may affect behavior, but is not taken as relevant to a social planner's evaluation. Potential examples of ancillary conditions include the point in time at which a choice must be made, the manner in which information or alternatives are presented, the labeling of a particular option as the "status-quo," the salience of a default option, or exposure to an anchor. Notably, if the individual's behavior appears to determine the ancillary condition endogenously, the decision problem has been defined incorrectly. For example, if he can choose to make his selection from X under one of two conditions, A or B, then it is inappropriate to describe A and B as endogenous ancillary conditions. Rather, the correct ancillary condition describes the two-stage decision process.

How does one objectively draw a line between the characteristics of the objects in the constraint set X and aspects of the ancillary condition d? In principle, one could view virtually any feature of a decision problem as a characteristic of the available objects. Yet if we incorporated *every* feature of each decision problem into the descriptions of the objects, then each object would be available in one and only one decision problem, and choices would provide little in the way of useful normative guidance. Consequently, practical considerations *compel* us to adopt a more limited conception of an object's attributes.

One natural way to draw the required line is to distinguish between conditions that pertain exclusively to experience and conditions that pertain at least in part to choice.

[27] Rubinstein and Salant [2008] independently formulated similar notation for describing the impact of choice procedures on decisions; they refer to ancillary conditions as "frames."

Conditions that pertain exclusively to experience do not change when a decision is delegated from an individual to a social planner. Consequently, if the planner treats such conditions as characteristics of the available objects, he can still take guidance from the choices the individual would make. If the planner must provide the individual with either a red car or a green car, he can sensibly ask which one the individual would choose; the meaning of color does not change with the chooser. In contrast, a condition that pertains to choice necessarily changes when the decision is delegated, because it then references a different chooser. If a planner were to treat such conditions as characteristics of the available objects, he would be forced to acknowledge that delegation necessarily changes the objects, in which case he would no longer be able to take guidance from a hypothetical undelegated choice. If he wishes to take such guidance, he must therefore define objects' characteristics to exclude conditions of choice. According to that criterion, hunger at the time of choice is an ancillary condition, but hunger at the time of consumption is not (a complete description of the good must specify consumption in each possible hunger state).

In many cases (e.g., when exposure to an arbitrary number influences choice), treating a condition of choice as a welfare-relevant characteristic of the available objects would seem to defy common sense; consequently, classifying it as an ancillary condition should be relatively uncontroversial. Other cases may be less clear. Different analysts may wish to draw different lines between the characteristics of choice objects and ancillary conditions, based either on the distinction between conditions of choice and experience discussed above, or using completely different criteria. It therefore bears emphasis that our framework provides a coherent method for conducting choice-based welfare analysis no matter how one draws that line. There is no escaping the fact that any application of our framework, like all other normative frameworks, must begin with some definition of the pertinent objects. However, critically, we are not compelled to use a definition that permits us to rationalize choices as reflecting well-behaved preferences. Where differences in line-drawing lead to different normative conclusions, our framework usefully pinpoints the source of disagreement.

Let $\mathcal{G}^*$ denote the set of all generalized choice situations contemplated by the positive theory of behavior for which we wish to develop a normative criterion. Thus, $\mathcal{G}^*$ is theory-specific. The positive theory implies a choice correspondence $C : \mathcal{G}^* \rightrightarrows \mathbb{X}$, with $C(X, d) \subseteq X$ for all $(X, d) \in \mathcal{G}^*$, that governs the individual's behavior.[28]

When confronted with conflicting choice patterns, behavioral economists sometimes argue that certain choices are more welfare-relevant than others. In effect, they prune elements of $\mathcal{G}^*$ from the welfare-relevant domain, so that the remaining choices coherently reveal "true" objectives. Often, such pruning is inadequately justified. Consider the example of time inconsistency. Suppose alternatives x and y yield payoffs at time t; the individual selects x over y when choosing at time t, and y over x when choosing at $t - 1$. Some argue that the individual's choice at $t - 1$ is the planner's best guide because it is at arms length from the experience and hence does not trigger the psychological processes responsible for apparent lapses of self-control; they would discard the time t choice. But one can also argue that the choice at t is the best guide because reward is properly appreciated only in the moment and excessively intellectualized at arm's length, which implies that one should discard the time $t - 1$ choice. Neither position is obviously superior. Retaining both choices and treating the time of choice as an ancillary condition permits us to recognize the conflict, remain agnostic, and embrace the implied ambiguity.

We allow for the possibility that pruning is adequately justified in some circumstances by defining a *welfare-relevant domain*, $\mathcal{G} \subseteq \mathcal{G}^*$, which identifies the choices from which we will take normative guidance. I will discuss potential *objective* criteria for pruning in Section 4.9. Meanwhile, my discussion will take $\mathcal{G}$ as given. Because our framework accommodates violations of standard choice axioms within $\mathcal{G}$, it permits one to demand more rigorous justifications for any deletions, even if the result is an enlarged domain that encompasses conflicting choices, such as $\mathcal{G}^*$.

[28] For our purposes, the nature of the evidence used to recover the choice correspondence is of no consequence. The reader is free to assume that positive analysis relies exclusively on choice evidence, or that non-choice evidence also plays a role.

Throughout this discussion, I will use $\mathcal{X}$ to denote the collection of constraint sets for which there is at least one ancillary condition d with $(X, d) \in \mathcal{G}$. For the most part, our analysis requires only two simple assumptions. First, $\mathcal{X}$ includes every finite subset of $\mathbb{X}$. Second, $C(G)$ is non-empty for all $G \in \mathcal{G}$.

Notice that our framework can incorporate non-standard behavioral patterns in four separate ways. (1) It allows choice to depend on ancillary conditions, thereby subsuming a wide range of behavioral phenomena. Specifically, it can accommodate any choice reversal involving a constraint set, X, along with two conditions, d' and d'', for which $C(X, d') \neq C(X, d'')$. (2) Our framework does not impose any counterparts to standard choice axioms. We allow for *all* non-empty choice correspondences, even ones for which choices are intransitive or depend on "irrelevant" alternatives (entirely apart from ancillary conditions). (3) Our framework subsumes the possibility that people can make choices from opportunity sets that are not compact (e.g., selecting "almost best" elements). (4) We can interpret a choice object $x \in \mathbb{X}$ more broadly than in the standard framework (e.g., as in Caplin and Leahy [2001], who axiomatize anticipatory utility by treating the time at which uncertainty is resolved as a characteristic of a lottery).

4.2 The welfare relations

Welfare analysis typically requires us to judge whether one alternative represents an *improvement* over another, even when the new alternative is not necessarily the best one. For that purpose, we require a binary relation, call it Q, where xQy means that x improves upon y. Within the standard framework, the revealed preference relations serve that role.

When the Weak Axiom of Revealed Preference (WARP) holds, one can define the weak and strict revealed preference relations in terms of choices from binary sets: xRy (xPy) iff $x \in C(\{x, y\})$ $(y \notin C(\{x, y\}))$. However, WARP also implies that one can equivalently define those relations in terms of choices from arbitrary sets, as follows:

$$xRy \text{ iff, for all } X \in \mathcal{X} \text{ with } x, y \in X,\ y \in C(X) \text{ implies } x \in C(X) \tag{4}$$

$$xPy \text{ iff, for all } X \in \mathcal{X} \text{ with } x, y \in X, \text{ we have } y \notin C(X) \tag{5}$$

Expressions (4) and (5) immediately suggest two natural generalizations of revealed preference. The first extends the notion of "no worse than" embedded in (4), the weak revealed preference relation:

$$xR'y \text{ iff, for all } (X,d) \in \mathcal{G} \text{ such that } x, y \in X,\ y \in C(X,d) \text{ implies } x \in C(X,d)$$

The statement "$xR'y$" means that whenever x and y are both available, y isn't chosen unless x is as well. We will then say that x is *weakly unambiguously chosen over y*. Let P' denote the asymmetric component of R' ($xP'y$ iff $xR'y$ and $\sim yR'x$), and let I' denote the symmetric component ($xI'y$ iff $xR'y$ and $yR'x$). The statement "$xP'y$" means that whenever x and y are available, sometimes x is chosen but not y, and otherwise either both or neither are chosen. The statement "$xI'y$" means that, whenever x is chosen, so is y, and vice versa.

The second generalization of revealed preference extends the notion of "better than" embedded in (5), strict revealed preference:

$$xP^*y \text{ iff, for all } (X,d) \in \mathcal{G} \text{ such that } x, y \in X, \text{ we have } y \notin C(X,d)$$

The statement "xP^*y" means that whenever x and y are available, y is never chosen. We then say that x is *strictly unambiguously chosen over y* (sometimes dropping "strictly" for the sake of brevity). As a general matter, P' and P^* may differ. However, if C maps each $G \in \mathcal{G}$ to a unique choice, they necessarily coincide.[29]

There are many binary relations for which P^* is the asymmetric component; each is a potential generalization of weak revealed preference. The coarsest is, of course, P^* itself. The finest, R^*, is defined by the property that xR^*y iff $\sim yP^*x$.[30] The statement "xR^*y" means that, for any $x, y \in \mathbb{X}$, there is *some* GCS for which x and y are available, and x is chosen; one can interpret it as signifying "might be better than." Let I^* be the symmetric

[29] Rubinstein and Salant [2008] have separately proposed a binary relation that is related to P' and P^*. See Bernheim and Rangel [2009] for a discussion of the relationship between our approaches.

[30] One binary relation A is *weakly coarser* than another, B, if xAy implies xBy. When A is weakly coarser than B, B is *weakly finer* than A.

component of R^* (xI^*y iff xR^*y and yR^*x). The statement "xI^*y" means that there is at least one GCS for which x is chosen with y available, and at least one GCS for which y is chosen with x available.

When choices are invariant with respect to ancillary conditions and satisfy standard axioms, R' and R^* specialize to R, while P' and P^* specialize to P. Thus, our framework subsumes standard welfare economics as a special case. More generally, it is easy to check that xP^*y implies $xP'y$ implies $xR'y$ implies xR^*y, so that P^* is the coarsest of these relations and R^* the finest. Also, $xI'y$ implies xI^*y.

The relation R^* is always complete, but R' need not be, and there is no guarantee that any of the relations defined here are transitive. However, to conduct useful welfare analysis, one does not require transitivity. Theorem 1 of Bernheim and Rangel [2009] establishes that there cannot be a cycle involving R', the natural generalization of weak revealed preference, if one or more of the comparisons involves P^*, the natural generalization of strict revealed preference. Thus, a planner who evaluates alternatives based on R' (to express "no worse than") and P^* (to express "better than") cannot be turned into a "money pump."[31] The theorem immediately implies that P^* is always acyclic. Like transitivity, acyclicity guarantees the existence of maximal elements for finite sets and allows us to both identify and measure unambiguous improvements. Thus, regardless of how poorly behaved the choice correspondence may be, P^* is always a viable welfare criterion.[32]

It is natural to wonder about the relationship between our approach and the multi-self criterion. Suppose in particular that $\mathcal{G}$ is the Cartesian product of the set of constraint sets and a set of ancillary conditions ($\mathcal{G} = \mathcal{X} \times D$, where $d \in D$); in that case, we say that $\mathcal{G}$ is *rectangular*. Suppose also that, for each $d \in D$, choices correspond to the maximal elements of a well-behaved preference ranking R_d, and hence to the alternatives that maximize a utility

[31] In the context of standard decision theory, Suzumura's [1976] analogous consistency property plays a similar role. A preference relation R is *consistent* if $x_1Rx_2...Rx_N$ with x_iPx_{i+1} for some i implies $\sim x_NRx_1$ (where P is the asymmetric component of R). If C maps each $G \in \mathcal{G}$ to a unique choice (so that P' coincides with P^*), then, according to this first theorem, R' is consistent.

[32] In contrast, it is easy to devise examples in which P' cycles.

function u_d. One can then imagine that each ancillary condition activates a different "self" and apply the Pareto criterion across selves. In that case, P^* is equivalent to strict multi-self Pareto domination, and P' to the weak version of that criterion (Bernheim and Rangel [2009], Theorem 3). Thus, in certain narrow settings, our approach justifies the multi-self Pareto criterion without invoking potentially controversial psychological assumptions, such as the existence of multiple coherent decision-making entities within the brain. Ironically, that justification does not apply to the standard model of quasi-hyperbolic discounting, which is the context in which the multi-self Pareto criterion is most commonly applied, because $\mathcal{G}^*$ is not rectangular. (I will elaborate on that point in Section 4.5 below, which explains the implications of our approach for quasi-hyperbolic decision makers.) However, it *does* justify the use of the multi-self Pareto criterion for cases of "coherent arbitrariness," such as those studied by Ariely, Loewenstein, and Prelec [2003], even though the criterion has not been applied in that context (see Section 4.5).

From the preceding comments, one should not form the incorrect impression that our framework fundamentally concerns the aggregation of conflicting preferences. Our welfare criterion is derived directly from choices and not from preferences, conflicting or otherwise; the fact that it sometimes coincides with a particular method of aggregating preferences for a specific as-if representation is merely an analytic convenience. To make the point more sharply (using the notation introduced in section 2.2), imagine that two decision process models, (I', P', γ') and (I'', P'', γ''), map to choice correspondences that coincide on the welfare-relevant domain. A welfare criterion based on preference aggregation might well generate different prescriptions for those two representations. In contrast, our welfare criterion is identical for any pair of observationally equivalent rationalizations because it depends only on the choice correspondence.

An advocate of the generalized revealed preference approach (discussed in Section 2) could potentially object to our welfare criterion on the grounds that, if a particular process model is known to be correct, and if its normative implications conflict with our criterion, then our framework necessarily generates false conclusions. Setting aside my skepticism concerning the prospects for developing the evidence necessary to support a normatively substantive process model, this objection ignores the fact that, within our framework, the same evidence would presumably justify restrictions on the welfare-relevant domain, $\mathcal{G}$, as well as (potentially) some reinterpretation of the choice correspondence.[33] I suspect that the correct application of the hypothesized evidence within our framework (as discussed in Section 4.9) would generally lead to normative conclusions that are consistent with the "true" decision process. For example, if it is known that an individual with well-behaved preferences sometimes "satisfices" when confronted with more than two alternatives, we would restrict $\mathcal{G}$ to binary choice sets, and thereby generate a welfare criterion that accurately reflects his well-being.

4.3 Welfare optima

We say that x is a *weak individual welfare optimum* in X if there is no $y \in X$ such that y improves upon x in the sense of P^*; formally, $\forall y \in X$, $\sim yP^*x$. Likewise, x is a *strict individual welfare optimum* in X if there is no $y \in X$ such that y improves upon x in the sense of P'; formally, $\forall y \in X$, $\sim yP'x$. Trivially, every $x \in C(X,d)$ for $(X,d) \in \mathcal{G}$ is a weak individual welfare optimum in X.[34] Thus, the existence of weak welfare optima is always guaranteed. Moreover, our welfare criterion respects a natural "libertarian" principle: any action voluntarily chosen from a set X within the welfare-relevant choice domain, $\mathcal{G}$, is a weak optimum within X. Thus, according to the relation P^*, it is impossible to design an intervention that "improves" on a choice made by the individual within $\mathcal{G}$. Nevertheless, it may be possible to improve decisions made in any GCS that is not considered welfare-relevant (i.e., elements of $\mathcal{G}^*$ that are excluded from $\mathcal{G}$, if any); see Section 4.9. It may also be possible to improve upon market outcomes when market failures are present, just as in

[33] For example, if the process model implies that the individual thinks of his opportunity set as X when in fact it is Y (either due to a misperception or an internal constraint), we would reinterpret a choice from Y as a choice from X.

[34] If x is the unique element of $C(X,d)$, then x is a strict welfare optimum in X.

standard economics; see Section 4.7.

The fact that the existence of weak individual welfare optima is guaranteed without any additional assumptions, e.g., related to continuity and compactness, may at first seem surprising, but simply reflects our assumption that the choice correspondence is well-defined over the set $\mathcal{G}$. Standard existence issues arise when the choice function is built up from other components. The following example clarifies these issues.

Example: Let $X_1 = \{a, b\}$, $X_2 = \{b, c\}$, $X_3 = \{a, c\}$, and $X_4 = \{a, b, c\}$, and suppose $\mathcal{G} = \{X_1, ..., X_4\}$ (plus singleton sets, for which choice is trivial). Notice that there are no ancillary conditions. Imagine that $C(X_1) = \{a\}$, $C(X_2) = \{b\}$, $C(X_3) = \{c\}$, and $C(X_4) = \{a\}$. Then we have aP^*b and bP^*c; in contrast, aI^*c. Despite the intransitivity of P^*, option a is nevertheless a strict welfare optimum in X_4, and neither b nor c is a weak welfare optimum. Note that a is also a strict welfare optimum in X_1 (b is not a weak optimum), and b is a strict welfare optimum in X_2 (c is not a weak optimum). Notably, both a and c are strict welfare optima in X_3, despite the fact that only c is chosen from X_3; a survives because it is chosen over c in X_4, which makes a and c not comparable under P^*.

Now let's limit attention to $\mathcal{G}' = \{X_1, X_2, X_3\}$. In that case, one of our underlying assumptions is violated ($\mathcal{G}'$ does not contain all finite sets) and P^* cycles ($aP^*bP^*cP^*a$). If we wish to create a preference or utility representation based on the data contained in $\mathcal{G}'$ so we can project the individual's choice within the set X_4, the intransitivity would pose a difficulty. And if we try to prescribe a welfare optimum for X_4 without knowing (either directly or through a positive model) what the individual would choose in X_4, we encounter the same problem: a, b, and c are all strictly improvable, so there is no welfare optimum. But as long as positive economists tell us what the individual would select from X_4, the existence problem for X_4 vanishes. □

As the previous example demonstrates, the alternatives chosen from a set need not be the only individual welfare optima within that set (specifically, a is an optimum in X_3, but is not chosen from X_3). As a general matter, x is a weak individual welfare optimum in

X if and only if for each $y \in X$ (other than x) there is some GCS for which x is chosen with y available (y may be chosen as well). The following example, based loosely on an experiment reported by Iyengar and Lepper [2000], illustrates why one can reasonably treat an alternative as an individual welfare optimum within a set even though the decision maker never chooses it from that set. Suppose a subject chooses strawberry jam when only one other flavor is available (regardless of what it is, and assuming he also has the option to take nothing), but rejects all flavors (including strawberry) in favor of nothing when thirty are available. In the latter case, one could argue that taking nothing is his best alternative because he chooses it. But one could also argue that strawberry jam is his best alternative because he chooses it over all of his other alternatives when facing simpler, less overwhelming decision problems. Our framework recognizes that both judgments are potentially valid on the basis of choice patterns alone.

4.4 A further justification for P^*

One could in principle formulate many binary relations based on choice; why P^*? Plainly, any weaker criterion would fail to respect an unambiguous choice pattern. Moreover, as I explain in this section, any stronger criterion would necessarily overrule choice (for some GCS within $\mathcal{G}$) by deeming an object improvable within a set from which it is chosen. In that important sense, it would violate the liberal principle that society should avoid second-guessing an individual's informed choice.

We say that a binary relation Q is an *inclusive libertarian relation* for a choice correspondence C if, for all X that appear in $\mathcal{G}$, the maximal elements under Q include all of the elements the individual would choose from X, considering all associated ancillary conditions. Such a relation never overrules choice, in the sense that any object chosen from a set X in some welfare-relevant condition is necessarily a weak individual welfare optimum within X. All other relations overrule choice in some circumstance. Thus, inclusive libertarianism is a desirable property for a choice-based welfare relation.

We have seen that P^* is an inclusive libertarian relation, but it is not the only one.

For example, the null relation, R^{Null} ($\sim xR^{Null}y$ for all $x, y \in \mathbb{X}$), falls into that category. Yet R^{Null} is less discerning than P^*. As it turns out, so are all other inclusive libertarian relations. In other words, P^* is always the most discerning inclusive libertarian relation. It follows that for all choice correspondences C and opportunity sets X, the set of maximal elements is weakly smaller under P^* than under any other inclusive libertarian relation (Bernheim and Rangel [2009], Theorem 2).

Ideally, for a given choice correspondence C, one might wish to find a binary welfare relation Q such that, for all $X \in \mathcal{X}$, the maximal elements under Q coincide *exactly* with the elements the individual would choose from X, considering all associated ancillary conditions. Such a relation would amount to a preference rationalization for choice. We know that, as a general matter, such rationalizations do not necessarily exist. However, according to the result in the previous paragraph, whenever there exists some preference relation that rationalizes choice on $\mathcal{G}$, P^* provides such a rationalization. In short, if there is a preference rationalization for choice, our framework employs that rationalization to evaluate welfare; when such a rationalization does not exist, our framework employs the most discerning relation that does not second-guess and overrule informed choice.

4.5 Applications to specific positive models

The Bernheim-Rangel framework is easily applied in the context of specific positive theories. Here, I discuss applications to two positive models: coherent arbitrariness and quasihyperbolic discounting.

Coherent arbitrariness. Behavior is coherently arbitrary when some psychological anchor (for example, calling attention to a number) affects choice, but the individual nevertheless conforms to standard axioms for any fixed anchor (see Ariely, Loewenstein, and Prelec [2003], who construed this pattern as an indictment of the revealed preference paradigm). To illustrate, let's suppose that an individual consumes two goods, y and z, and that we have the following representation of decision utility:

$$U(y, z \mid d) = u(y) + dv(z)$$

with u and v strictly increasing, differentiable, and strictly concave. We interpret the ancillary condition, $d \in [d_L, d_H]$, as an anchor that influences decision utility.

Because $\mathcal{G}$ is rectangular, and because choices maximize $U(y, z \mid d)$ for each d, our welfare criterion is equivalent to the multi-self Pareto criterion, where each d indexes a different self (Theorem 3 of Bernheim and Rangel [2008]). It follows that

$$(y', z')R'(y'', z'') \text{ iff } u(y') + dv(z') \geq u(y'') + dv(z'') \text{ for } d = d_L, d_H \tag{6}$$

Replacing the weak inequality with a strict one, we obtain a similar equivalence for P^*.

Figure 1(a) shows two decision-indifference curves (that is, indifference curves derived from decision utility) passing through the bundle (y', z'), one for d_L (labelled I_L) and one for d_H (labelled I_H). All bundles (y'', z'') lying below both decision-indifference curves satisfy $(y', z')P^*(y'', z'')$; this is the analog of a lower contour set. All bundles (y'', z'') lying above both decision-indifference curves satisfy $(y'', z'')P^*(y', z')$; this is the analog of an upper contour set. For all bundles (y'', z'') lying between the two decision-indifference curves, we have *neither* $(y', z')R'(y'', z'')$ nor $(y'', z'')R'(y', z')$; however, $(y', z')I^*(y'', z'')$.

Now consider a standard budget constraint, $X = \{(y, z) \mid y + pz \leq M\}$, where y is the numeraire, p is the price of z, and M is income. As shown in Figure 1(b), the individual chooses bundle a when the ancillary condition is d_H, and bundle b when the ancillary condition is d_L. Each of the points on the thick segment of the budget line between bundles a and b is uniquely chosen for some $d \in [d_L, d_H]$, so all these bundles are strict individual welfare optima. It is easy to prove that there are no other welfare optima, weak or strict.

As the gap between d_L and d_H shrinks, the set $\{(y'', z'') \mid (y'', z'')P^*(y', z')\}$ converges to a standard upper contour set, and the set of individual welfare optima converges to a single utility maximizing choice. Thus, our welfare criterion converges to a standard criterion as the behavioral anomaly becomes small. As I discuss below, that observation reflects a general principle.

Dynamic inconsistency. Economists who use β, δ model of quasihyperbolic discounting for policy analysis tend to employ one of two welfare criteria: either the multi-self Pareto criterion, which associates each moment in time with a different self, or the "long-run criterion," which assumes that well-being is described by exponential discounting at the rate δ. Our framework leads to a different criterion.

As in Section 2.2, suppose the consumer's task is to choose a consumption vector, c. Choices at time t maximize the function

$$U_t(C_t) = u(c_t) + \beta \sum_{k=t+1}^{T} \delta^{k-t} u(c_k) \ , \tag{7}$$

where $\beta, \delta \in (0, 1)$. We assume perfect foresight concerning future decisions, so that behavior is governed by subgame perfect equilibria. We also assume $u(0)$ is finite; for convenience, we normalize $u(0) = 0$. Finally, we assume $\lim_{c \to \infty} u(c) = \infty$.

To conduct normative analysis, we must recognize that the selection of an intertemporal consumption vector involves only one choice by a single decision maker. Critically, that statement remains valid even when the individual makes the decision over time in a series of steps (notwithstanding the common practice of modeling such problems as games between multiple time-dated selves); he still selects a single consumption trajectory. For this positive model, a GCS $G = (X, \tau)$ involves a set of lifetime consumption vectors, X, and a decision tree, τ, for selecting an element of X (as discussed in Section 2.2). For every possible constraint set X, $\mathcal{G}^*$ includes every conceivable pair (X, τ), where τ is decision tree for selecting from X. Note that $\mathcal{G}^*$ is not rectangular: decision trees are tailored to constraint sets, and in any case the individual cannot chose consumption for period t using a tree that allows no choice until period $k > t$. Hence, our sufficient conditions for justifying the multi-self Pareto criterion do not apply.

Assume for now that the welfare-relevant domain $\mathcal{G}$ coincides with the full choice domain

$\mathcal{G}^*$. Define $W(c) \equiv \sum_{k=t}^{T} (\beta\delta)^{k-t} u(c_k)$. It turns out that $c' R' c''$ iff $W(c') \geq U_1(c'')$, and $c' P^* c''$ iff $W_1(c') > U_1(c'')$ (Bernheim and Rangel [2009], Theorem 4). This result tells us, in effect, that it is possible to design an intrapersonal game in which c'' is chosen when c' is feasible if and only if $W(c') < U_1(c'')$. Thus, to determine whether c' is unambiguously chosen over c'', we compare the first period decision utility obtained from c'' (that is, $U_1(c'')$) with the first period utility obtained from c' discounting at the rate $\beta\delta$ (that is, $W(c')$). Given our normalization ($u(0) = 0$), we necessarily have $U_1(c') \geq W(c')$, from which two conclusions follow. First, R' and P^* are transitive. Second, $U_1(c') > U_1(c'')$ is a necessary (but not sufficient) condition for c' to be unambiguously chosen over c''; in other words, any welfare improvement under P^* or P' must also be a welfare improvement under U_1, the decision utility at the first moment in time. Our characterization result also implies that c is a weak welfare optimum in X if and only if the decision utility that c provides at $t = 1$ is at least as large as the highest available discounted value of u, using $\beta\delta$ as a time-consistent discount factor (formally, iff $U_1(c) \geq \sup_{c' \in X} W(c')$).

The characterization result has the surprising implication that $c'_t > c''_t$ for all t does not necessarily imply $c' P^* c''$. The reason is that one can design intrapersonal games in which the individual chooses dominated consumption vectors. For example, we might offer him a choice between c' and c'' in period 1, but if he chooses c', allow him to swap c' for some c''' in period 2. By making c''' sufficiently attractive (in terms of decision utility) from the period 2 perspective but sufficiently unattractive from the period 1 perspective, we can induce him to choose c'' in period 1. If one is uncomfortable with the normative implication that $\sim c' P^* c''$, that discomfort simply reflects an implicit judgment that the choice situation in question is not normatively relevant (e.g., because it reflects the inefficient interplay between conflicting objectives). By making that judgment explicit, we can generate a refinement for the β, δ model (see Section 4.9).

Notice that, for all c, $\lim_{\beta \to 1} [W(c) - U_1(c)] = 0$. Accordingly, as the degree of dynamic inconsistency shrinks, our welfare criterion converges to the standard criterion. In contrast, the same statement does *not* hold for the multi-self Pareto criterion, as that criterion is usually formulated. The reason is that, regardless of β, each self is assumed to care only about current and future consumption. Thus, consuming everything in the final period is always a multi-self Pareto optimum, even when $\beta = 1$.

Note that if the relevant time periods are short (e.g., days) and the value of β is noticeably less than one (e.g., 0.95), then the welfare criterion identified in Theorem 4 may be discerning only when applied to problems with short planning horizons (e.g., short-term procrastination, but not retirement). In Section 4.9, I discuss potential criteria for restricting the welfare-relevant domain $\mathcal{G}$, thereby generating more discerning criteria.

4.6 Tools for applied welfare analysis

The concepts of compensating and equivalent variation have natural counterparts within our framework. To illustrate, let's assume that the individual's constraint set, $X(\alpha, m)$, depends on a vector of environmental parameters, α, and a monetary transfer, m. Let α_0 be the initial parameter vector, d_0 the initial ancillary condition, and $(X(\alpha_0, 0), d_0)$ the initial GCS. We will consider a change in parameters to α_1 and in the ancillary condition to d_1, along with a monetary transfer m. We write the new GCS as $(X(\alpha_1, m), d_1)$. This setting will allow us to evaluate compensating variations for fixed changes in prices, ancillary conditions, or both.[35]

In seeking to generalize the notion of compensating variation, we immediately encounter an important ambiguity concerning the standard of compensation: do we consider compensation sufficient when the new situation (with the compensation) is unambiguously chosen over the old one, or when the old situation is not unambiguously chosen over the new one? That ambiguity is an essential feature of welfare evaluations with inconsistent choice. Accordingly, we define two notions of compensating variation:

[35] This formulation of compensating variation assumes $\mathcal{G}$ is rectangular. If $\mathcal{G}$ is not rectangular, then as a general matter we would need to write the final GCS as $(X(\alpha_1, m), d_1(m))$, and specify the manner in which d_1 varies with m.

Definition: CV-A is the level of compensation m^A that solves

$$\inf\{m \mid yP^*x \text{ for all } m' \geq m,\, x \in C(X(\alpha_0, 0), d_0) \text{ and } y \in C(X(\alpha_1, m'), d_1)\}$$

Definition: CV-B is the level of compensation m^B that solves

$$\sup\{m \mid xP^*y \text{ for all } m' \leq m,\, x \in C(X(\alpha_0, 0), d_0) \text{ and } y \in C(X(\alpha_1, m'), d_1)\}$$

In other words, all levels of compensation greater than the CV-A (smaller than CV-B) guarantee that everything selected in the new (initial) set is unambiguously chosen over everything selected from the initial (new) set. It is easy to verify that $m^A \geq m^B$. Also, when $\alpha_1 = \alpha_0$ and $d_1 \neq d_0$, we always have $m^A \geq 0 \geq m^B$. Thus, the welfare effect of a change in the ancillary condition, by itself, is always ambiguous.

CV-A and CV-B are well-behaved welfare measures in the following sense: If the individual experiences a sequence of changes and is adequately compensated for each in the sense of the CV-A, no alternative he would select from the initial set is unambiguously chosen over any alternative he would select from the final set. Similarly, if he experiences a sequence of changes and is not adequately compensated for any of them in the sense of the CV-B, no alternative he would select from the final set is unambiguously chosen over any alternative he would select from the initial set.

Under more restrictive assumptions, the compensating variation of a price change corresponds to an analog of consumer surplus. Consider again the model of coherent arbitrariness, but assume a more restrictive form of decision utility (which involves no income effects, so that Marshallian consumer surplus would be valid in the standard framework):

$$U(y, z \mid d) = y + dv(z) \tag{8}$$

Thus, for any given d, the inverse demand curve for z is given by $p = dv'(z) \equiv P(z, d)$.

Let M denote the consumer's initial income. Consider a change in the price of z from p_0 to p_1, along with a change in ancillary conditions from d_0 to d_1. Let z_0 denote the amount of z purchased with (p_0, d_0), and let z_1 denote the amount purchased with (p_1, d_1); assume that $z_0 > z_1$. Define $m(d) = [p_1 - p_0]z_1 + \int_{z_1}^{z_0}[P(z,d) - p_0]dz$. Then $m^A = m(d_H)$ and $m^B = m(d_L)$ (Bernheim and Rangel [2009], Theorem 5). The first term in the expression for $m(d)$ is the extra amount the consumer pays for the first z_1 units. The second term involves the area between the demand curve and a horizontal line at p_0 between z_1 and z_0 when d is the ancillary condition.

Figure 2(a) provides a graphical illustration of CV-A, analogous to ones found in most microeconomics textbooks: it is the sum of the areas labeled A and B. Figure 2(b) illustrates CV-B: it is the sum of the areas labeled A and C, minus the area labeled E. Note that CV-A and CV-B bracket the conventional measure of consumer surplus that one would obtain using the demand curve associated with the ancillary condition d_0. As the range of possible ancillary conditions narrows, CV-A and CV-B both converge to standard consumer surplus. As I discuss below in Section 4.8, that property reflects a general principle.

For an application of this framework to a practical problem involving the salience of sales taxes, as well as for an extension to settings with income effects, see Chetty et. al. [2008].

4.7 Welfare analysis involving more than one individual

Suppose there are N individuals indexed $i = 1, ..., N$. We now construe $\mathbb{X}$ as the set of all conceivable social alternatives. Let C_i be the choice correspondence for individual i, defined over $\mathcal{G}_i$ (where the subscript reflects the possibility that the set of ancillary conditions may differ from individual to individual). These choice correspondences induce the relations R_i' and P_i^*over $\mathbb{X}$. We say that x is a *weak generalized Pareto optimum* in X if there exists no $y \in X$ with yP_i^*x for all i. We say that x is a *strict generalized Pareto optimum* in X if there exists no $y \in X$ with $yR_i'x$ for all i, and yP_i^*x for some i. If one thinks of P^* as a preference relation, then our notion of a weak generalized Pareto optimum coincides with existing notions of social efficiency when consumers have incomplete and/or intransitive preferences (see, e.g., Fon and Otani [1979], Rigotti and Shannon [2005], or Mandler [2006]).

Since strict individual welfare optima do not always exist, we cannot guarantee the existence of strict generalized Pareto optima with a high degree of generality. However, we can trivially guarantee the existence of a weak generalized Pareto optimum for any set X: simply choose $x \in C_i(X, d)$ for some i and $(X, d) \in \mathcal{G}_i$.

In the standard framework, there is typically a continuum of Pareto optima that spans the gap between the extreme cases in which the chosen alternative is optimal for some individual. We often represent that continuum by drawing a utility possibility frontier or, in the case of a two-person exchange economy, a contract curve. Is there also usually a continuum of generalized Pareto optima spanning the gap between the extreme cases described in the previous paragraph? Just as in the standard framework, one can start with *any* alternative $x \in X$ and find a Pareto optimum $y \in X$ such that no individual unambiguously chooses x over y (Bernheim and Rangel [2009], Theorem 6). Consequently, as the following example illustrates, utility possibility frontiers and contract curves have natural counterparts.

Example: Consider a two-person exchange economy involving two goods, y and z. Suppose the choices of consumer 1 are described by the model of coherent arbitrariness discussed earlier, while consumer 2's choices respect standard axioms. In Figure 3, we have drawn two standard contract curves. The one labeled T_H is formed by the tangencies between the consumers' indifference curves when consumer 1 faces ancillary condition d_H (such as the point at which I_{1H} touches I_2); the one labeled T_L is formed by the tangencies when consumer 1 faces ancillary condition d_L (such as the point at which I_{1L} and I_2). The shaded area between those two curves is the generalized contract curve; it contains all the weak generalized Pareto optimal allocations. The ambiguities in consumer 1's choices *expand* the set of Pareto optima, which is why the generalized contract curve is thick.[36] Like a standard contract curve, the generalized contract curve runs between the southwest and northeast corners of the Edgeworth box, so there are many intermediate Pareto optima. If the behavioral effects of the ancillary conditions were smaller, the generalized contract curve would be thinner; in the limit, it would converge to a standard contract curve. (As discussed

[36] Mandler [2008] demonstrates with generality that the Pareto efficient set has full dimensionality.

below in Section 4.8, this point reflects a general principle.) □

The notion of a generalized Pareto optimum easily lends itself to formal analysis. As an illustration, let's examine the efficiency of competitive equilibria. Consider an economy with N consumers, F firms, and K goods. We place no restrictions on consumers' choice correspondences, aside from non-emptiness. Firms are endowed with production technologies and are assumed to maximize profits. Define a *behavioral competitive equilibrium* as a price vector and a vector of ancillary conditions that clear all markets.[37] In Bernheim and Rangel [2009], we demonstrate that, if all choices are welfare-relevant ($\mathcal{G}_n = \mathcal{G}_n^*$), then the allocation associated with any behavioral competitive equilibrium is a weak generalized Pareto optimum (Theorem 7).[38]

The generality of this result is worth emphasizing: it establishes the efficiency of competitive equilibria within a framework that imposes almost no restrictions on consumer behavior, thereby allowing for virtually any conceivable choice pattern, including all anomalies documented in the behavioral literature. Note, however, that the theorem need not hold if firms pursue objectives other than profit maximization. Thus, we see that the first welfare theorem is driven by assumptions concerning the behavior of firms, not consumers.

Naturally, a behavioral competitive equilibrium can be inefficient in the presence of sufficiently severe but otherwise standard market failures. In addition, even a perfectly competitive behavioral equilibrium may be inefficient when judged by a welfare relation derived from a restricted welfare-relevant choice domain ($\mathcal{G}_n \subset \mathcal{G}_n^*$). This observation alerts us to the fact that, in behavioral economics, there is a new class of potential market failures involving choice situations that have been pruned from $\mathcal{G}_n^*$. Our analysis of addiction (Bernheim and Rangel [2004]) exemplifies that possibility.

[37] One could endogenize the ancillary conditions by supplementing this definition with additional equilibrium requirements. However, our result would still apply.

[38] Our analysis extends a result due to Fon and Otani [1979] involving economies wherein consumers have incomplete and/or intransitive preferences; see also Rigotti and Shannon [2005] and Mandler [2006]

4.8 Standard welfare analysis as a limiting case

Several of the examples in the preceding sections suggest that, for settings in which departures from standard choice axioms are minor, one can approximate the appropriate welfare criterion by ignoring choice anomalies and applying the standard normative framework. In Bernheim and Rangel [2009], we establish that point with generality, providing formal convergence results for upper and lower contour sets (Theorem 8), compensating variation (Theorem 9), weak Pareto optima (Theorem 10), and individual welfare optima (Corollary 3). The additional assumptions required to establish these results are mild.

Our analysis of convergence is important for three reasons. First, it justifies the common view that the standard welfare framework must be approximately correct when behavioral anomalies are small. A formal justification for that view has been absent. To conclude that the standard normative criterion is roughly correct in a setting with choice anomalies, we would need to compare it to the correct criterion. Unless we have established the correct criteria for such settings, we have no benchmark against which to gauge the performance of the standard criterion, even when choice anomalies are tiny. Our framework overcomes that problem by providing welfare criteria for all situations. Our results imply that small choice anomalies have only minor implications for welfare. Thus, we have formalized the intuition that a little bit of positive falsification is unimportant from a *normative* perspective.

Second, our convergence results imply that the debate over the significance of choice anomalies need not be resolved prior to adopting a framework for welfare analysis. If our framework is adopted and the anomalies ultimately prove to be small, one will obtain virtually the same answer as with the standard framework.

Third, our convergence results suggest that our welfare criterion will always be reasonably discerning provided behavioral anomalies are not too large. That observation is reassuring, in that the welfare relations may be extremely coarse, and the sets of individual welfare optima extremely large, when choice conflicts are sufficiently severe.

4.9 Refining the welfare relations

It is straightforward to verify that R' and P^* become weakly finer as the welfare-relevant domain ($\mathcal{G}$) shrinks and weakly coarser as it expands. Intuitively, if choices between two alternatives x and y are unambiguous over some domain, they are also unambiguous over a smaller domain, but they may be ambiguous over a larger domain. Consequently, if one is concerned that R' and P^* are insufficiently discerning, one can potentially refine those relations by excluding GCSs from the welfare-relevant domain. Justifying such refinements generally requires one to officiate between apparent choice conflicts. Many existing discussions of behavioral welfare economics amount to informal arguments concerning officiation; for example, one choice is sometimes taken to be more indicative of "true normative preferences" than another (e.g., Bernheim and Rangel [2004], Noor [2008a,b]). Our framework permits one to introduce and formalize such arguments within the context of identifying $\mathcal{G}$.

For a choice-based normative framework, it is natural to consider the possibility of self-officiating through meta-choices (that is, choices between choices). The case of time-inconsistency illustrates some of the conceptual problems with that approach. Assume an individual would choose x over y for time t at time t, but would choose y over x for time t at time $t-1$. Any meta-choice between those choices must occur at time $t-1$ or earlier. Therefore, just like the decision at $t-1$, all meta-choices are made at arms length from the reward. But an arms-length choice clearly cannot objectively resolve whether another arms-length choice (the one at time $t-1$) or an in-the-moment choice (the one at time t) is a more appropriate normative guide.

More generally, a meta-choice is simply another GCS. Within our framework, consideration of meta-choices therefore amounts to expanding the welfare-relevant domain $\mathcal{G}$, which makes the relations R' and P^* weakly coarser, potentially enlarging (and never shrinking) the set of weak individual welfare optima. The welfare relations can become finer only if we also exclude the "defeated" GCS, which would implicitly require us to elevate the status of one type of choice (the meta-choice) over another (the original choice). But that elevated status necessarily reflects an arbitrary judgment. We might seek a choice-based justification for that judgment by considering a second-level meta-choice (between the original meta-choice and the excluded GCS), but that path leads inevitably to consideration of higher and higher level meta-choices, with no logical stopping point. Also, unless one is willing to impose additional structure, there is no guarantee that meta-choices will be decisive; for example, they may be cyclic, or k-th level meta-choices may conflict with $(k+1)$-th level meta-choices for all k. Thus, it is hard to imagine a compelling choice-based justification for deference to meta-choices.

Can we devise other compelling criteria for excluding GCSs from the welfare-relevant domain, $\mathcal{G}$? The remainder of this section discusses several possibilities.

Refinements based on imperfect processing of factual information. Suppose there is some GCS, $G = (X, d)$, in which the individual incorrectly perceives the constraint set as $Y \neq X$. It is then appropriate to delete that GCSs from the welfare-relevant domain $\mathcal{G}$.[39] Even with its deletion, ambiguities in R' and P^* may remain, but those relations nevertheless become (weakly) finer and hence more discerning.

Why would the individual believe himself to be choosing from the wrong set? His attention may focus on some small subset of X, his memory may fail to call up facts that relate choices to consequences, he may forecast the consequences of his choices incorrectly, or he may have learned from his past experiences more slowly than the objective information would permit. Accordingly, we propose using non-choice evidence, including findings from psychology, neuroscience, and neuroeconomics, to identify and delete *suspect* GCSs in which those types of informational processing failures occur. We potentially classify the choice in a suspect GCS as a “mistake,” not because we second-guess the desirability of the outcome, but rather because we have found direct objective evidence that the decision maker did not fully grasp or attend to the problem.

[39] In principle, if we understood the individual’s cognitive processes sufficiently well, we might be able to identify his perceived choice set Y, and reinterpret the choice as pertaining to Y rather than to X. While it may be possible to accomplish that task in some instances (see, e.g., Koszegi and Rabin [2008b]), I suspect that, in most cases, it is beyond the current capabilities of economics, neuroscience, and psychology.

The following simple example motivates the use of evidence from neuroscience. An individual is offered a choice between alternatives x and y. He chooses x when the alternatives are described verbally, and y when they are described partly verbally and partly in writing. Which choice is the best guide for public policy? If we learn that the information was provided in a dark room, we would be inclined to respect the choice of x, rather than the choice of y. We would reach the same conclusion if an opthamologist certified that the individual was blind, or, more interestingly, if a brain scan revealed that his visual processing circuitry was impaired. In all these cases, non-choice evidence sheds light on the likelihood that the individual successfully processed information that was in principle available to him, thereby properly identifying the choice set X.

Our work on addiction (Bernheim and Rangel [2004]) illustrates this agenda. Citing evidence from neuroscience, we argue as follows. First, the brain's value forecasting circuitry includes a specific neural system that measures empirical correlations between cues and potential rewards. Second, the repeated use of an addictive substance causes that system to malfunction in the presence of cues that are associated with its use. Whether or not that system *also* plays a role in hedonic experience, the choices made in the presence of those cues are therefore predicated on improperly processed factual information, and welfare evaluations should be guided by choices made under other conditions (e.g., precommitments).

In many situations, simpler forms of evidence may suffice. For example, if an individual consistently characterizes a choice as a mistake on the grounds that he neglected or misunderstood factual information, or if a simple test of his knowledge reveals that he ignored critical information, then one might justifiably declare the choice suspect, at least provisionally until further evidence is assembled. Other considerations, such as the complexity of a GCS, could also come into play.

It is important to emphasize a critical difference between the uses of non-choice evidence that I have endorsed in this section and those that I criticized in Section 2.2. Conceptually, decision making involves two distinct tasks. The first, *characterization*, involves identifying the opportunities available to the individual, as well as the mapping from actions to consequences. The second, *choice*, involves the selection of an alternative from the perceived opportunity set. The skepticism I expressed in Section 2.2 concerns the prospects for understanding the second process (choice) well enough to justify a single normatively unambiguous model of decision making. I am much more optimistic about the prospects for identifying *characterization failures* involving errors in processing factual information, in part because the issues are mostly mechanical rather than interpretive, and in part because the task does not require a comprehensive understanding of the full decision process.

Refinements based on coherence. In some instances, it may be possible to partition behavior into coherent patterns and isolated anomalies. One might then adopt the position that welfare analysis should ignore the isolated anomalies entirely. That argument suggests another potential refinement strategy: identify subsets of GCSs within which choices are coherent (in the sense that standard axioms hold); then construct welfare relations based on those GCSs and ignore other choices. The main difficulty with this *coherence criterion* is that all behavior is coherent within a sufficiently narrow scope (e.g., every choice is coherent taken by itself). How does one judge whether that scope is too narrow? Despite our inability to offer a general and precise definition, there are nevertheless contexts in which coherence has a natural interpretation.

Take the problem of intertemporal consumption allocation for a β, δ consumer (Section 4.5). Consider a *single-choice GCS* (in which the decision is completely resolved through full precommitment at a single point in time) that conflicts with a *staged-choice GCS* (in which it is made in a series of steps). Much of the pertinent literature adopts the position that a decision for the single-choice GCS reflects a single coherent perspective while a decision for the staged-choice GCS does not. Indeed, as noted in Section 4.5, an individual can select a dominated consumption vector in a staged-choice GCS. Those considerations invite an application of the coherence criterion: exclude staged-choice GCSs from the welfare-relevant domain, denoted $\mathcal{G}_c$, while retaining single-choice GCSs, which cohere within time-indexed subsets. Let R'_c and P^*_c denote the resulting welfare relations.

Because the proposed refinement does not officiate between conflicting single-choice GCSs, it fails to resolve all ambiguity. Nevertheless, within our framework, it yields a discerning welfare criterion. Specifically, taking $\mathcal{G}_c$ as the welfare-relevant domain, a consumption trajectory is a welfare optimum within any standard intertemporal budget constraint if and only if it maximizes decision utility at the first moment in time, $U_1(c)$ (Bernheim and Rangel [2009], Theorem 11). According to that result, welfare optimality within such constraint sets is completely governed by the individual's perspective at the first moment in time. The special status of $t = 1$, which I noted in Section 4.5, is amplified when attention is restricted to $\mathcal{G}_c$. Thus, even though the coherence criterion does not resolve all choice conflicts, it justifies the judgements embedded in the long-run criterion (exponential discounting at the rate δ) for certain environments, assuming the first period is short.

Bernheim and Rangel [2009] also demonstrate that the coherence criterion, as applied to quasihyperbolic decision makers, isolates precisely the same set of individual welfare optima as robust multi-self Pareto optimality (a concept mentioned Section 2.2). Intuitively, if the welfare-relevant domain were rectangular, P^*_c would coincide with the strict multi-self Pareto relation (as discussed in Section 4.2). We can make it rectangular by hypothetically extending the choice correspondence C to include choices involving past consumption. Deleting those hypothetical choices makes the welfare relation more discerning and does not enlarge the set of weak individual welfare optima. Thus, the set of weak individual welfare optima under P^*_c must lie within the set of multi-self Pareto optima for every conceivable pattern of backward-looking choices. In light of this result, the special status of the $t = 1$ perspective is intuitive: that perspective dominates robust multi-self Pareto comparisons because we lack critical information (backward-looking preferences) concerning all other perspectives.

Refinements based on other criteria. If people process information more completely and accurately when making straightforward choices, a *simplicity criterion* could have merit. That criterion would presumably favor one-shot binary decision problems. Unfortunately, if we construct P^* exclusively from data on binary decisions, acyclicity is not guaranteed (recall my first example). However, in certain settings, that procedure does generate coherent welfare relations. Consider again the β, δ model of quasihyperbolic discounting. Fixing the date of choice at time t, behavior within the set of one-shot binary decision problems fully "reveals" the decision-utility function U_t, as does behavior within the set of single-choice GCSs. Therefore, officiating in favor of one-shot binary decision problems is equivalent to officiating in favor of single-choice GCSs; both approaches lead to the welfare relations R_c' and P_c^*.

One could also apply a *preponderance criterion*: if someone ordinarily chooses x over y and rarely chooses y over x, disregard the exceptions and follow the rule. That criterion is sometimes invoked (at least implicitly) in the literature on quasi-hyperbolic (β,δ) discounting to justify use of the long-run perspective: trade-offs between rewards in periods t and $t+k$ are governed only by δ from the perspective of all periods $s < t$, and by both β and δ only from the perspective of period t.

I see two conceptual problems with the preponderance criterion. First, there are potentially many competing notions of frequency. Because it is possible to proliferate variations of ancillary conditions, one cannot simply count GCSs. In the quasi-hyperbolic setting, a count of time-dated perspectives would favor the long-run criterion. However, an application of preponderance based on the frequency with which GCSs are encountered (an index of familiarity) might favor the short-run perspective.

Second, a rare ancillary condition may be highly conducive to good decision-making. That would be the case, for example, if an individual typically misunderstands available information concerning his alternatives unless it is presented in a particular way. Likewise, in the quasi-hyperbolic setting, one could argue that people may appreciate their needs most accurately when those needs are immediate and concrete, rather than distant and abstract.

5 Conclusions

How one approaches the task of generalizing standard welfare analysis to settings with nonstandard decision makers depends on one's interpretation of the standard framework. According to one interpretation, standard normative analysis respects the decision maker's true objectives, which her choices reveal. Because the behaviors of interest by definition defy conventional rationalizations, that interpretation requires one to entertain unconventional rationalizations. But as a general matter one can offer many unconventional rationalizations for any particular behavioral pattern. Thus, knowledge of a choice correspondence may shed insufficient light on objectives, and hence on the mapping from the objects of choice to well-being. One can attempt to identify welfare either partially or completely by restricting the set of allowable unconventional rationalizations, but useful restrictions are difficult to justify. Those conceptual difficulties have led some to argue that economists should try to infer well-being from self-reported happiness and/or neurobiological activity. Unfortunately, it is every bit as problematic to identify useful information concerning internal well-being from such data as it is from choice.

Fortunately, there is an attractive alternative. According to a second interpretation of standard welfare analysis, welfare is *defined* in terms of choice rather than underlying objectives. As my work with Antonio Rangel demonstrates (Bernheim and Rangel [2007, 2008, 2009]), that interpretation leads to a rich and tractable normative framework. Our approach exploits coherent aspects of choice by replacing the standard revealed preference relation with an *unambiguous choice* relation. I summarized the attractive features of our framework at the outset of Section 4, and will refrain from repeating myself here.

In closing, I will briefly mention some alternative perspectives that deviate more radically from conventional welfare economics. Sugden [2004] defines welfare in terms of opportunity, and conducts formal normative analysis based on an "opportunity criterion." See also Sen [1992] and Arrow [1995], who argue that opportunity has intrinsic value, and Roemer [1998], who suggests that public policy should concern itself with opportunity, rather than with how people use opportunities. Sen [1985] proposes defining welfare in terms of intermediate states between consumption and utility, which he terms "functionings." For example, nourishment is a functioning that lies between the consumption of food and feelings of happiness and satisfaction. Sen speculates that there may be widespread agreement on the rankings of many functioning vectors, even if people disagree about the rankings of consumption vectors, and he maintains that functionings are interpersonally comparable. Sen [1977] argues for a normative framework based on "meta-rankings." He hypothesizes that a person who sometimes acts on different preference orderings in different contexts would have preferences over those preference orderings. He argues that such meta-rankings should guide normative analysis because they reflect higher ethical principles. Gul and Pesendorfer [2008] advocate defining welfare in terms of political sustainability. For example, they argue that Pareto optimality is a sensible welfare criterion because a Pareto inefficient institution would be unstable. Due to space constraints, a comprehensive discussion of the advantages and limitations of those proposals lies beyond the scope of the current paper.

I will close with a final word concerning the relationship between the ideas discussed in this paper and the concept of *libertarian paternalism*, which has recently attracted some attention. In proposing that concept, Thaler and Sunstein [2003] argued that choices are sensitive to objectively irrelevant conditions (such as the manner in which information is presented); therefore, any party who is in a position to influence those conditions, either by action or inaction, cannot avoid a degree of paternalism. Libertarian paternalism calls for minimal restrictions on individual discretion, but recognizes the inevitable impact of a planner's decisions on the conditions of individual choice; it calls upon the planner to establish conditions that tend to produce good decisions. The challenge, of course, is to distinguish objectively between good and bad outcomes. Thaler and Sunstein offer a few suggestions as well as some context-specific illustrations in which choices are transparently mistaken. Their suggestions fall within the categories discussed in this paper: in some cases, they would take selective guidance from choices (as I discussed in Sections 2.2 and 4.9); in others, they would employ *ex post* measures of well-being (as I discussed in Section 3). They do not, however, propose a general welfare criterion or normative framework.

References

[1] Ariely, Dan, George Loewenstein, and Drazen Prelec. 2003. "Coherent Arbitrariness: Stable Demand Curves without Stable Preferences." *Quarterly Journal of Economics*, 118(1):73-105.

[2] Arrow, Kenneth J. 1959. "Rational Choice Functions and Orderings." *Economics*, 26(102): 121-127.

[3] Arrow, Kenneth J. 1995. "A Note on Freedom and Flexibility." In Kaushik Basu, Prasanta Pattanaik, and Kotaro Suzumura (eds.), *Choice,welfare and development: A festschrift in honour of Amartya K. Sen.* Oxford: Oxford University Press.

[4] Asheim, Geir. 2008. "Procrastination, Partial Naivete, and Behavioral Welfare Analysis." Mimeo, University of Oslo.

[5] Bernheim, B. Douglas. 1999. "Comment on 'Family Bargaining and Retirement Behavior."' In Henry Aaron (ed.), *Behavioral Economics and Retirement Policy.* Washington, D.C.: Brookings Institution Press.

[6] Bernheim, B. Douglas. 2009. "Neuroeconomics: A Sober (but Hopeful) Appraisal." *AEJ Microeconomics*, forthcoming.

[7] Bernheim, B. Douglas, and Antonio Rangel. 2004. "Addiction and Cue-Triggered Decision Processes." *American Economic Review,* 94(5):1558-90.

[8] Bernheim, B. Douglas, and Antonio Rangel. 2007. "Toward Choice-Theoretic Foundations for Behavioral Welfare Economics." *American Economic Review Papers and Proceedings* 97(2), 464-470.

[9] Bernheim, B. Douglas, and Antonio Rangel. 2008. "Choice-Theoretic Foundations for Behavioral Welfare Economics." In Andrew Caplin and Andrew Schotter (eds.), *The Methodologies of Modern Economics.* Oxford University Press, forthcoming.

[10] Bernheim, B. Douglas, and Antonio Rangel. 2009. “Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics.” *Quarterly Journal of Economics*, forthcoming.

[11] Caplin, Andrew, and John Leahy. 2001. “Psychological Expected Utility Theory and Anticipatory Feelings.” *The Quarterly Journal of Economics,* 116(1): 55-79.

[12] Carmichael, H. Lorne, and W. Bentley MacLeod. 2006. “Welfare Economics with Intransitive Revealed Preferences: A Theory of the Endowment Effect.” *Journal of Public Economic Theory*, 8 (2): 193–218

[13] Chambers, Christopher, and Takashi Hayashi. 2008. “Choice and Individual Welfare.” Mimeo, California Institute of Technology.

[14] Cherepanov, V., T. Feddersen, and A. Sandroni. 2008. “Rationalization.” Mimeo, Northwestern University.

[15] Chetty, Raj, Adam Looney, and Kory Kroft. 2008. “Salience and Taxation: Theory and Evidence." Mimeo, University of California, Berkeley.

[16] Dalton, Patricio, and Sayantan Ghosal. 2008. “Behavioral Decisions and Welfare.” Mimeo, University of Warwick.

[17] Dekel, Eddie, and Barton Lipman. 2007. “Self-Control and Random Strotz Representations.” Mimeo, Harvard University.

[18] Easterlin, Richard A., 1974 . “Does Economic Growth Improve the Human Lot?: Some Empirical Evidence.” In P. A. David and W. R. Levin (eds.), *Nations and Household in Economic Growth.* Stanford University Press.

[19] Easterlin, Richard A., 1995. “Will Raising the Incomes of All Increase the Happiness of All?” *Journal of Economic Behavior and Organization*, 27: 35-47.

[20] Easterlin, Richard A., 2003. "Explaining Happiness," *Proceedings of the National Academy of Science*, 100 (19): 11176-11183.

[21] Fon, Vincy, and Yoshihiko Otani. 1979. "Classical Welfare Theorems with Non-Transitive and Non-Complete Preferences." *Journal of Economic Theory*, 20: 409-418.

[22] Frey, Bruno S., and Alois Stutzer. 2002. "What Can Economists Learn from Happiness Research?" *Journal of Economic Literature* 40: 402-435.

[23] Frey, Bruno S., and Alois Stutzer. 2004. "Economic Consequences of Mispredicting Utility." Institute for Empirical Research in Economics Working Paper #218, University of Zurich.

[24] Green, Jerry, and Daniel Hojman. 2007. "Choice, Rationality, and Welfare Measurement." Mimeo, Harvard University.

[25] Gruber, Jonathan, and Sendhil Mullainathan. 2005. "Do Cigarette Taxes Make Smokers Happier?" *Advances in Economics Analysis & Policy*, Berkeley Electronic Press,

[26] Gul, Faruk, and Wolfgang Pesendorfer. 2001. "Temptation and Self-Control." *Econometrica,* 69(6):1403-1435.

[27] Gul, Faruk, and Wolfgang Pesendorfer. 2008. "The Case for Mindless Economics." In Andrew Caplin and Andrew Schotter (eds.), *The Methodologies of Modern Economics.* Oxford University Press, forthcoming.

[28] Iyengar, S. S., and M. R. Lepper. 2000. "Why Choice is Demotivating: Can One Desire Too Much of a Good Thing?" *Journal of Personality and Social Psychology* 79, 995-1006.

[29] Kahneman, D. 1999. "Objective Happiness." In Kahneman, D., E. Diener, and N. Schwarz (eds.), *Well-Being: The Foundations of Hedonic Psychology.* New York: Russell Sage Foundation.

[30] Kahneman, Daniel and Robert Sugden. 2005. “Experienced Utility as a Standard of Policy Analysis.” *Environmental and Resource Economics* 32(1): 161-81

[31] Kahneman, D., P. Wakker, and R. Sarin. 1997. “Back to Bentham? Explorations of Experienced Utility.” *Quarterly Journal of Economics*, 112: 375-406.

[32] Kimball, Miles, Helen Levy, Fumio Ohtake and Yoshiro Tsutsui. 2006. “Unhappiness after Hurricane Katrina,” NBER Working Paper #12062.

[33] Kimball, Miles and Robert Willis. 2006. “Utility and Happiness,” mimeo, University of Michigan.

[34] Koszegi, Botond, and Matthew Rabin. 2008a. “Choices, Situations, and Happiness.” *Journal of Public Economics*, forthcoming.

[35] Koszegi, Botond, and Matthew Rabin. 2008b. “Revealed Mistakes and Revealed Preferences.” In Andrew Caplin and Andrew Schotter (eds.), *The Methodologies of Modern Economics.* Oxford University Press, forthcoming.

[36] Krusell, Per, Burhanettin Kuruscu, Anthony A. Smith. 2001. “Temptation and Taxation.” Mimeo, Princeton University.

[37] Laibson, David. 1997. “Golden Eggs and Hyperbolic Discounting.” *Quarterly Journal of Economics*, 112(2):443-477

[38] Laibson, David, Andrea Repetto, and Jeremy Tobacman. 1998. “Self-Control and Saving for Retirement.” *Brookings Papers on Economic Activity*, 1: 91-172.

[39] Larsen, Randy J., and Barbara L. Fredrickson. 1999. “Measurement Issues in Emotion Research.” In Daniel Kahneman, Ed Diener and Norbert Schwarz (eds.), *Well-Being: The Foundations of Hedonic Psychology.* New York: Russell Sage Foundation.

[40] Layard, Richard. 2005. *Happiness.* London, UK: Penguin Press.

[41] Little, I. M. D. 1949. "A Reformulation of the Theory of Consumer's Behaviour." *Oxford Economic Papers*, 1(1): 90-99.

[42] Loewenstein, George; O'Donoghue, Ted; Rabin, Matthew. 2003. "Projection Bias in Predicting Future Utility." *Quarterly Journal of Economics*, 118 (4): 1209-48.

[43] Mandler, Michael. 2006. "Welfare Economics with Status Quo Bias: A Policy Paralysis Problem and Cure." Mimeo, University of London.

[44] Mandler, Michael. 2008. "Indecisiveness in Behavioral Welfare Economics." Mimeo, University of London.

[45] Mandler, Michael, Paola Manzini, and Marco Mariotti. 2008. "A Million Answers to Twenty Questions: Choosing by Checklist." Mimeo, University of London.

[46] Manzini, Paola, and Marco Mariotti. 2008. "Categorize Then Choose: Boundedly Rational Choice and Welfare." Mimeo, University of London.

[47] McClure, S. M., D. I. Laibson, G. Loewenstein, and J. D. Cohen. 2004. "Separate neural systems value immediate and delayed monetary rewards." *Science* 306: 503–507.

[48] Mill, John Stuart. 1869. *On Liberty.* London: Longman, Roberts, & Green.

[49] Noor, Jawwad. 2008a. "Subjective Welfare." Mimeo, Boston University.

[50] Noor, Jawwad. 2008b. "Temptation, Welfare, and Revealed Preference." Mimeo, Boston University.

[51] Nozick, Robert. 1974. *Anarchy, State, and Utopia.* New York: Basic Books.

[52] O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing It Now or Later." *American Economic Review*, 89(1):103-24.

[53] Rayo, Luis, and Gary Becker. 2007. "Habits, Peers, and Happiness: An Evolutionary Perspective." *American Economic Review*, 97(2): 487-91.

[54] Rigotti, Luca, and Chris Shannon. 2005. "Uncertainty and Risk in Financial Markets." *Econometrica*, 73(1): 203-243.

[55] Roemer, Jobn E. 1998. *Equality of opportunity*. Cambridge, MA: Harvard University Press.

[56] Rubinstein, Ariel, and Yuval Salant. 2008. "(A,f) Choice with frames." *Review of Economic Studies*, forthcoming.

[57] Samuelson, Paul. 1938. "A Note on the Pure Theory of Consumer's Behaviour." *Economica*, 5(17): 61-71

[58] Sen, Amartya K. 1973. "Behavior and the Concept of Preference." *Economica* 40: 241-59.

[59] Sen, Amartya K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy and Public Affairs*, 6(4): 317-44.

[60] Sen, Amartya K. 1985. *Commodities and Capabilities*. Amsterdam & New York: North-Holland.

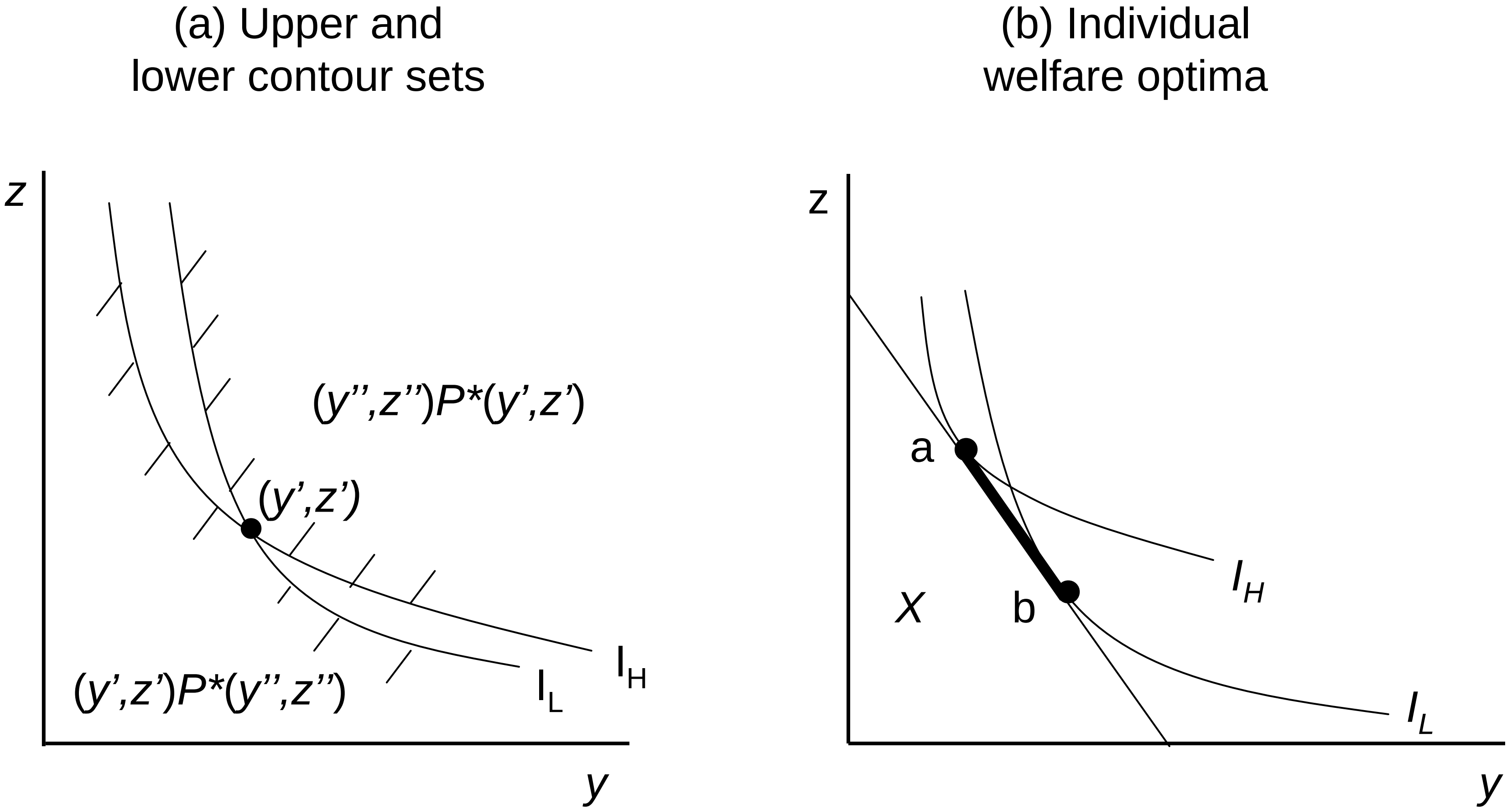
[61] Sen, Amartya K. 1992. *Inequality reexamined*. Cambridge, MA: Harvard University Press.

[62] Simon, H.A. 1976. "From Substantive to Procedural Rationality." In S. J. Latsis (ed.), *Methods and Appraisal in Economics*. Cambridge University Press.

[63] Sugden, Robert. 2004. "The Opportunity Criterion: Consumer Sovereignty Without the Assumption of Coherent Preferences." *American Economic Review*, 94(4): 1014-33.

[64] Suzumura, Kotaro. 1976. "Remarks on the Theory of Collective Choice." *Economica*, 43: 381-390.

[65] Thaler, Richard, and Cass R. Sunstein. 2003. “Libertarian Paternalism.” *American Economic Review Papers and Proceedings*, 93(2): 175-179.



(a) Upper and lower contour sets

(b) Individual welfare optima

Figure 1: Coherent arbitrariness

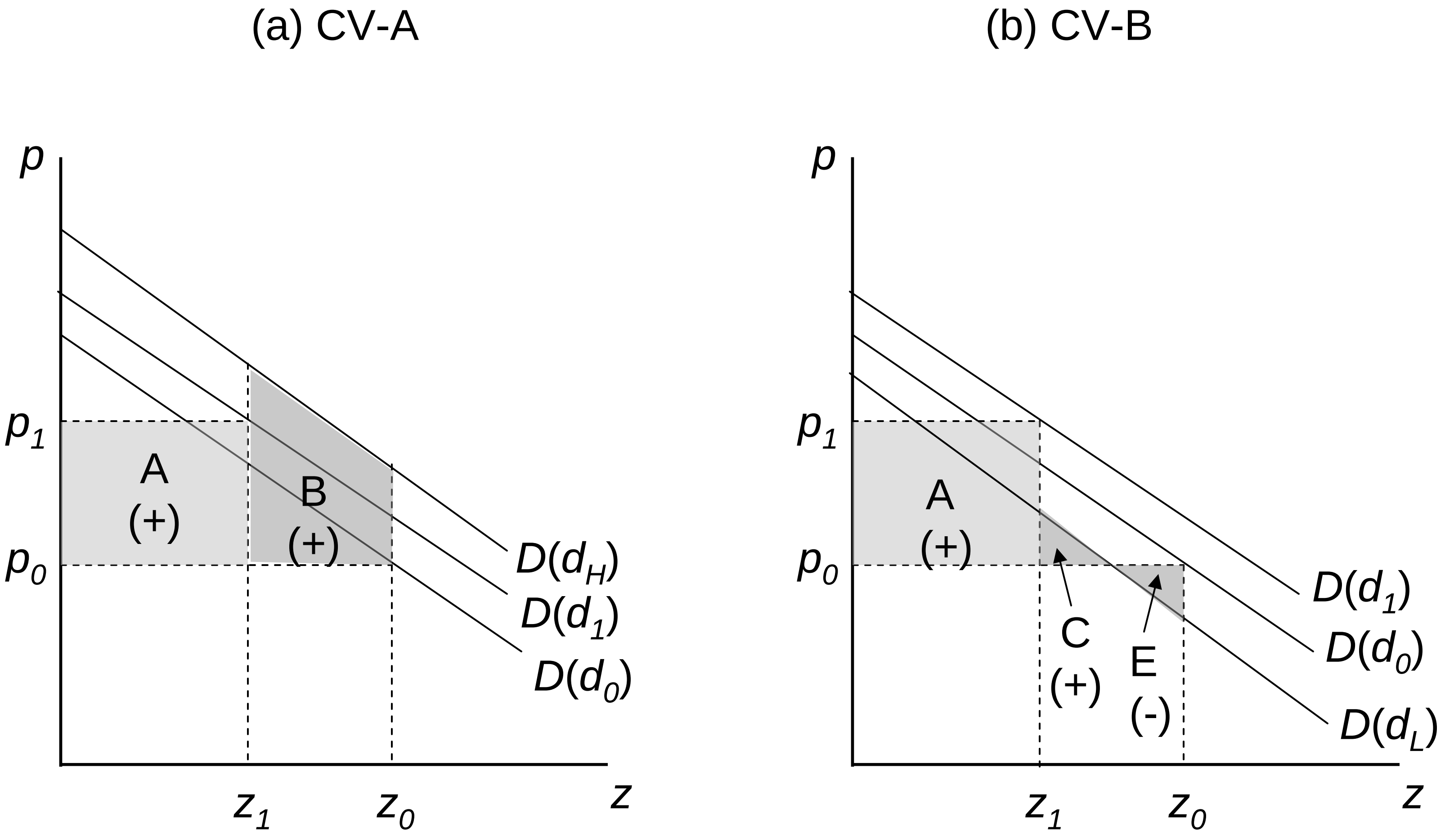


Figure 2: CV-A and CV-B for a change from (p_0,d_0) to (p_1,d_1)

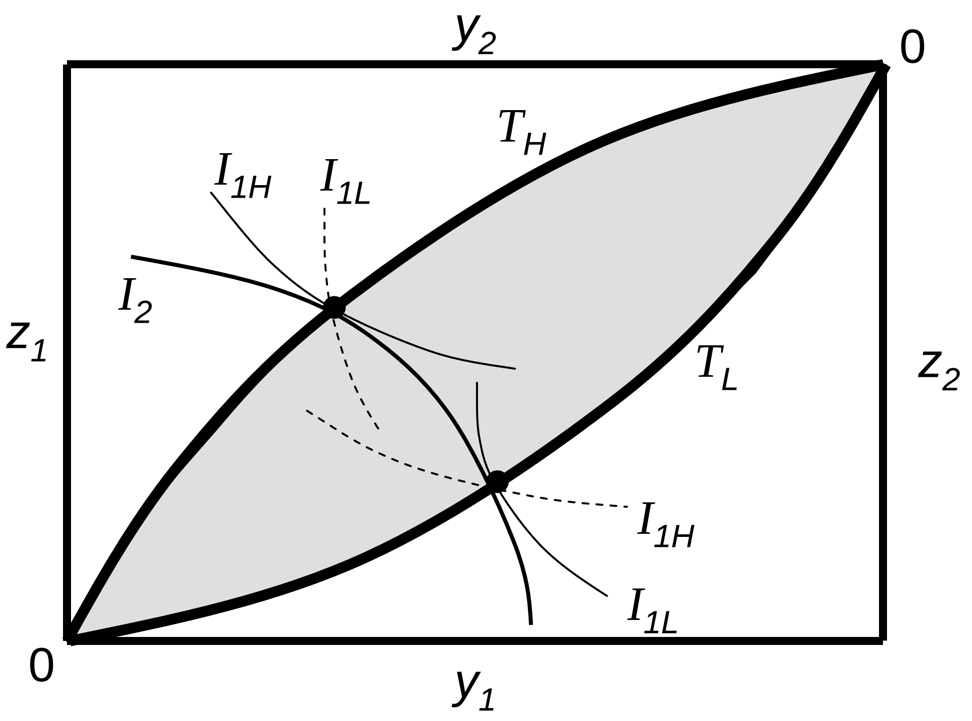


Figure 3: The generalized contact curve