Copyright Notice

Bernheim, B. Douglas, and Antonio Rangel. 2007. “Toward Choice-Theoretic Foundations for Behavioral Welfare Economics.” American Economic Review 97 (2): 464–70. https://doi.org/10.1257/aer.97.2.464.

Toward Choice-Theoretic Foundations for Behavioral Welfare Economics

By B. Douglas Bernheim and Antonio Rangel*

Interest in behavioral economics has grown in recent years, stimulated largely by accumulating evidence that the standard model of consumer decision making provides an inadequate, positive description of human behavior. Behavioral models are increasingly finding their way into policy evaluation, which inevitably involves welfare analysis. No consensus concerning the appropriate standards and criteria for behavioral welfare analysis has emerged yet.

This paper summarizes our effort to develop a unified framework for behavioral welfare economics (for a detailed discussion see Bernheim and Rangel 2007)—one that can be viewed as a natural extension of standard welfare economics. Standard welfare analysis is based on choice, not on utility or preferences. In its simplest form, it instructs the planner to respect the choices an individual would make for himself. The guiding normative principle is an extension of the libertarian deference to freedom of choice, which takes the view that it is better to give a person the thing he would choose for himself rather than something that someone else would choose for him.

We show that it is possible to extend the standard choice-theoretic approach to welfare analysis to situations where individuals make inconsistent choices, which are prevalent in behavioral economics.

[†] *Discussant:* Colin Camerer, California Institute of Technology.

* Bernheim: Department of Economics, Stanford University, Stanford, CA 94305 (e-mail: bernheim@stanford.edu); Rangel: Division of Humanities and Social Science, California Institute of Technology, 228-77, Pasadena, CA 91125 (e-mail: rangel@hss.caltech.edu).

I. Preliminaries

A. *Choice Situations and Choices*

We use $\mathbb{X}$ to denote the set of all possible choice objects. Elements of $\mathbb{X}$ could be standard objects such as state-contingent consumption bundles or lotteries over such bundles, but these objects could also have nonstandard features (e.g., as in Andrew Caplin and John Leahy 2001).

A *standard choice situation* (SCS) consists of a constraint set $X \subseteq \mathbb{X}$, representing the set of objects from which the individual can choose according to the objective information at his disposal. We are concerned with making welfare judgments based on behavior for some particular domain of standard choice situations, $\mathcal{X}$.

To accommodate choice inconsistencies, we introduce the notion of a *generalized choice situation* (GCS), G. This consists of a standard choice situation, X, and a set of ancillary conditions, d. Thus, $G = (X,d)$. We will use $\mathcal{G}$ to denote the set of generalized choice situations of potential interest. Examples of ancillary conditions include the point in time at which the decision is made, the manner in which information is presented, and the labelling of some options as defaults. Behavior is represented by a choice correspondence $C : \mathcal{G} \to \mathbb{X}$; for any $G \in \mathcal{G}$, any $x \in C(G)$ is an action that the individual is willing to choose.

As a general matter, it is difficult to draw a bright line between the characteristics of the objects in X and the ancillary conditions d (which, in principle, one could also view as a characteristic of the objects in the choice set). Our analytic framework is applicable regardless of how one draws this distinction.

Standard economics proceeds from the assumption that choice is invariant with respect

to ancillary conditions. Positive behavioral economics challenges this basic premise since there are many examples of *choice reversals* in which $C(X,d') \neq C(X,d'')$ for two ancillary conditions d' and d''.

B. *Positive versus Normative Analysis*

Usually, choice data are not available for all elements of $\mathcal{G}$, but rather for elements of some set $\mathcal{H} \subset \mathcal{G}$. The objective of positive economic analysis is to extend the choice correspondence C from observations on to the entire set $\mathcal{G}$. In standard economics, this is accomplished by defining preferences over $\mathbb{X}$, estimating these preferences with choice data for the opportunity sets in $\mathcal{H}$, and using these estimated preferences to infer choices for opportunity sets in $\mathcal{G}\backslash\mathcal{H}$.

The objective of normative economic analysis is to evaluate outcomes. Typically, we evaluate outcomes at the level of the individual, and then aggregate. For choice-based normative criteria, we allow the individual's choices to govern these evaluations. The fundamental problem of behavioral welfare economics is to identify appropriate criteria for evaluating alternatives when, due to choice reversals and other behavioral anomalies, the individual's choices fail to provide clear guidance.

In conducting choice-based normative analysis, we take as given the individual's choice correspondence, C, defined on $\mathcal{G}$ rather than $\mathcal{H}$. Preferences and utility functions, which are constructs used both in standard theory and in behavioral economics to extend C from $\mathcal{H}$ to $\mathcal{G}$, are therefore positive tools, not normative tools. In a behavioral setting, these constructs cannot meaningfully reconcile inconsistencies. They can only reiterate the information contained in the extended choice correspondence C. Thus, one cannot resolve normative puzzles by identifying classes of preferences that rationalize apparently inconsistent choices, as in Faruk Gul and Wolfgang Pesendorfer (2001).

II. A Framework for Behavioral Welfare Analysis

In the standard approach to normative analysis, one evaluates individual welfare by applying a "revealed preference" relation, R, defined over the elements of $\mathbb{X}$, which summarizes what is chosen from various SCSs. Under standard assumptions, R is an ordering. When we use this ordering to conduct welfare analysis, we are simply asking what an individual or individuals would choose. For example, the compensating variation associated with some change in the economic environment equals the smallest payment that would induce the individual to choose the change. Similarly, a social alternative $x \in X$ is a Pareto optimum in X if there is no other alternative in X that all individuals would choose over x.

In behavioral economics, we often cannot summarize all choices with a consistent preference ordering. Instead, choice evidence is sometimes ambiguous. Fortunately, we can live with this ambiguity. It is still possible to construct binary relations based on *unambiguous* comparisons that allow us to carry out meaningful welfare analyses.

A. *Individual Welfare Orderings*

In standard welfare economics, the statement xRy means that if x and y are available in X, and if y is in $C(X)$, then x is also in $C(X)$. Our proposal is to conduct behavioral welfare economics by generalizing this binary relation. In effect, xRy will mean that if x and y are available in G, and if y is in $C(G)$, then x is in $C(G)$. We do not pretend that this relation reveals preference. It is simply a summary of what is chosen.

More specifically, for any $x,y \in \mathbb{X}$, we will say that xRy if, whenever x and y are available, x is sometimes chosen, and y is never chosen unless x is as well. Formally, (i) there exists some $(X,d) \in \mathcal{G}$ with $\{x,y\} \subseteq X$ such that $x \in C(X,d)$, and (ii) there does not exist any $(X,d) \in \mathcal{G}$ with $\{x,y\} \in X$ such that $y \in C(X,d)$ and $x \notin C(X,d)$. Note that xRx.

As usual, we can define xPy as xRy and $\sim yRx$. This means that, whenever x and y are available, sometimes x is chosen but not y, and otherwise either both or neither is chosen. Likewise, we can define xIy as xRy and yRx. This means that, whenever x is chosen, so is y, and vice versa.

Finally, we will say that xP^*y if, whenever x and y are available, sometimes x is chosen but not y, and otherwise neither is chosen. Formally, (i) for all $(X,d) \in \mathcal{G}$ with $\{x,y\} \subseteq X$, we have

$y \notin C(X,d)$, and (ii) for some $(X,d) \in \mathcal{G}$ with $\{x,y\} \subseteq X$, we have $x \in C(X,d)$.

In general, R, P, and P^* need not be orderings. For example, if $C(\{x,y\},d') = \{x\}$ and $C(\{x,y\},d'') = \{y\}$, then we have *neither* xRy *nor* yRx, so R is not complete. Moreover, R, P, and P^* need not be transitive (though P^* is acyclic). For example, if choice does not depend on ancillary conditions, and if we have $C(\{x_1,x_2\}) = x_1$, $C(\{x_2,x_3\}) = x_2$, $C(\{x_3,x_1\}) = x_3$, and $C(\{x_1,x_2,x_3\}) = \{x_1,x_2,x_3\}$, then $x_1 P x_2 P x_3 P x_1$.

B. *Individual Welfare Optima*

There are two natural criteria for determining whether a choice is improvable. We will say that it is possible to *strictly improve* upon a choice $x \in X$ if there exists $y \in X$, such that yP^*x. In other words, there is an alternative that is unambiguously chosen over x. When a strict improvement is impossible, we say that x is a *weak individual welfare optimum*. It is possible to weakly *improve* upon a choice $x \in X$ if there exists $y \in X$ such that yPx. In other words, there is an alternative y that is sometimes chosen over x, and that x is never chosen over y (except in the sense that both could be chosen). When a weak improvement is impossible, we say that x is a *strict individual welfare optimum*.

When is $x \in X$ an individual welfare optimum? The following simple results (which we state without proof) assume that $\mathcal{G}$ includes all pairwise comparisons (that is, for all $a,b \in \mathbb{X}$, there is some d_{ab} such that $(\{a,b,d_{ab}) \in \mathcal{G}$).

FACT 1: *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 present (y may be chosen as well).*

This result has an immediate corollary. If x is chosen for some GSC involving X, then x is a weak individual welfare optimum in X. Notice that this corollary guarantees the existence of weak individual welfare optima.

FACT 2: *x is a strict individual welfare optimum in X if for each $y \in X$ (other than x) either x is chosen and y is not for some GCS with y present, or there's no GCS for which y is chosen and x is not with x present.*

This result also has an immediate corollary: if x is the unique choice for some GSC involving X, then x is a strict individual welfare optimum in X. A strict individual welfare optimum may not exist (see the example given at the end of Section IA).

C. *Relationship to Multi-Self Pareto Optima*

Our notion of an individual welfare optimum is related to the idea of a multi-self Pareto optimum. Suppose, in particular, that the set of GCSs is the Cartesian product of the set of SCSs and a set of ancillary conditions (that is, $\mathcal{G} = \mathcal{X} \times D$, where $d \in D$). Also imagine that, for each $d \in D$, choices follow standard axioms, and can be represented by a preference ranking R_d. If one imagines that each ancillary condition activates a different "self," then one can conduct welfare analysis by examining multi-self Pareto optima. Under the stated conditions, a weak multi-self Pareto optimum corresponds to a weak individual welfare optimum (as we have defined it), and a strict multi-self Pareto optimum corresponds to a strict individual welfare optimum. For these narrow settings, our approach is equivalent to the multi-self Pareto criterion. Our approach is also more general, however, in that it does not require the assumptions stated at the outset of this paragraph. Moreover, it can justify the multi-self Pareto criterion without reference to questionable psychological assumptions.

D. *Equivalent and Compensating Variation*

The concepts of equivalent and compensating variation are central to applied welfare economics. Fortunately, they have natural counterparts within our framework. Here, we will focus on compensating variation. Our treatment of equivalent variation is analogous.

We will write the individual's SCS as $X(\alpha,m)$, where α is a vector of parameters, and m is the level of compensation. Let α_0 be the initial parameter vector, d_0 the initial ancillary conditions, and $(X(\alpha_0,0),d_0)$ the initial GCS. We will consider a change in parameters to α_1, coupled with some level of compensation, as well as (potentially)

a change in ancillary conditions that could, in principle, depend on the compensation level. We write the new GCS as $(X(\alpha_1,m),d(m))$. One natural possibility, but certainly not the only one, is to take $d(m) = d'$ for some fixed d'. This allows us, for example, to evaluate compensating variations for fixed changes in prices, ancillary conditions, or both.

Here, we will define the compensating variation relative to particular selections from the choice correspondence (see Bernheim and Rangel 2007 for some alternatives). Accordingly, we assume the individual selects $x_0 \in C(X(\alpha_0,0),d_0)$, and $x(m) \in C(X(\alpha_1,m),d(m))$.

In defining the compensating variation, we encounter an ambiguity concerning the standard of compensation. Do we consider compensation sufficient when $x(m)$ is always weakly chosen over x_0, or when x_0 is not always weakly chosen over $x(m)$? This ambiguity is an essential feature of welfare evaluations with inconsistent choice. Accordingly, we define two notions of compensating variation:

$$m^{CV-A} = \inf_m x(m) P^* x_0,$$
$$m^{CV-B} = \sup_m x_0 P^* x(m).$$

We illustrate the application of these concepts by discussing the measurement of consumer surplus.

E. *Consumer Surplus*

For simplicity, we will examine a case where the individual consumes two goods: x and y. Suppose that positive analysis delivers the following utility representation (which involves no income effects, so that Marshallian consumer surplus would be valid in the standard framework):

$$U(x,y|d) = x + dv(y),$$

with v strictly increasing, differentiable, and strictly concave. Thus, for any given d, the inverse demand curve for y is given by $p = dv'(y)$, where p is the price of y. Notice that the ancillary condition, $d \in [d_L,d_H]$, simply shifts the weight attached to $v(y)$. This might, for example, represent the type of "coherent arbitrariness" documented by Dan Ariely, George Loewenstein, and Drazen Prelec (2003).

Let M denote the consumer's initial income. Consider a change in the price of y from p_0 to p_1, along with a change in ancillary conditions from d_0 to d_1. Let y_0 denote the amount of y purchased at price p_0, and let y_1 denote the amount purchased at price p_1. Assume that $y_0 > y_1$. Since there are no income effects, y_1 will not change as the individual is compensated (holding the ancillary condition fixed).

Now let us calculate CV-A. We wish to find the smallest m such that $(y_1, M - p_1y_1 + m)$ $P^*(y_0, M - p_0y_0)$. It is straightforward to show that the solution is

$$m^{CV-A} = [p_1 - p_0]y_1 + \int_{y_1}^{y_0} [d_H v'(y) - p_0]\ dy.$$

Through similar reasoning, one can show that

$$m^{CV-B} = [p_1 - p_0]y_1 + \int_{y_1}^{y_0} [d_L v'(y) - p_0]\ dy.$$

Figures 1(A) and 1(B) illustrate m^{CV-A} (the shaded area above p_0) and m^{CV-B} (the shaded area above p_0 minus the shaded area below p_0), respectively, for the case where $d_0 = d_1 = d$. Notice that these values bracket standard consumer surplus. Moreover, as the range of possible ancillary conditions narrows, they converge to standard consumer surplus. This underscores the fact that the standard framework is a special case of the framework considered here. Moreover, it also implies that, when inconsistencies are minor (that is, $d_H - d_L$ is small), the ambiguity in welfare, as measured by the difference between m^{CV-A} and m^{CV-B}, is small.

F. *Generalized Pareto Optima*

Next, we turn to environments with many individuals, and formulate a generalization of Pareto efficiency. Suppose there are N individuals indexed $i = 1,\ldots,N$. Let $\mathbb{X}$ denote the set of all conceivable social choice objects, and let X denote the set of feasible objects. Let C_i be the choice function for individual i, defined over $\mathcal{G}_i$ (where the subscript reflects the possibility that

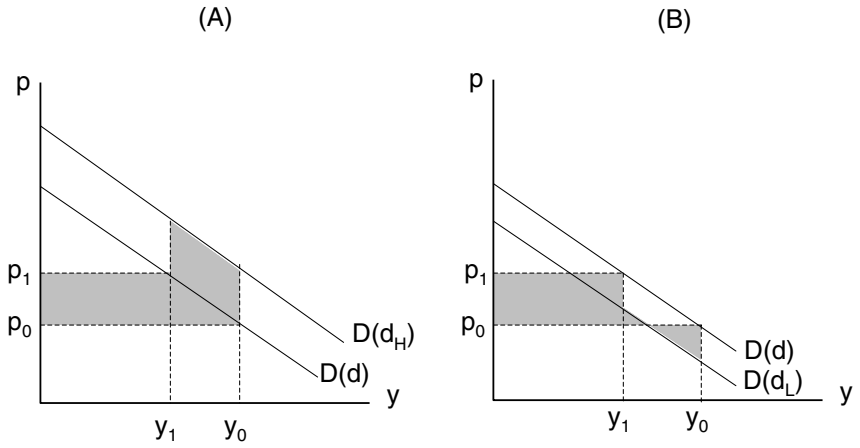

FIGURE 1

the set of ancillary conditions may differ from individual to individual). These choice functions induce the relations R_i and P_i^* over $\mathbb{X}$.

We will 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 will say that x is a *strict generalized Pareto optimum* in X if there exists no $y \in X$ with yR_ix for all i, and yP_i^*x for some i.

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. We *can* guarantee the existence of a weak generalized Pareto optimum, however.

G. *The Efficiency of Competitive Equilibria*

To illustrate the usefulness of these concepts, we have provided a generalization of the first welfare theorem. Specifically, we consider a production economy consisting of N consumers, F firms, and K goods. The economy is standard in all respects, except that consumer i's behavior is governed by a general choice correspondence mapping budget sets and ancillary conditions into sets of consumption vectors. We make one simple assumption (akin to nonsatiation) with respect to consumer behavior: if $x^n > w^n$ (where $>$ indicates a strict inequality for every good), then consumer n never chooses w^n when x^n is available.

A *behavioral competitive equilibrium* involves a price vector, $\hat{\pi} = (\hat{\pi}^1,\ldots,\hat{\pi}^K)$, along with a vector of ancillary conditions, $\hat{d} = (\hat{d}^1,\ldots,\hat{d}^N)$, which clear all markets. Though behavioral competitive equilibria may not exist, those that do exist are necessarily efficient.

THEOREM 1: *The allocation in any behavioral competitive equilibrium is a strict generalized Pareto optimum.*

It is worth emphasizing that a perfectly competitive equilibrium may be inefficient when judged by a refined welfare relation, after officiating choice conflicts, as described in the next section. This observation alerts us to the fact that, in behavioral economics, choice reversals lead to a new class of potential market failures.

III. Refinements

A. *The Logic of Refinements*

In any particular context, the relation R and P^* that we have defined may not be very discerning, which means that many choice alternatives might be individual welfare optima. In this section, we consider the possibility that one might refine the relations R and P^* by altering the data used to construct them. Most obviously, one can add choice data (by creating new GCSs, expanding the domain $\mathcal{G}$), or delete data (by ignoring certain GCSs, reducing the domain $\mathcal{G}$). There is also the possibility of reinterpreting choice data, which we mention briefly below.

We say that R' is *coarser* than R if $xR'y$ implies xRy. When R' is coarser than R, we say that R is *finer* than R'. Subject to a technical qualification, the addition of data (that is, the expansion of $\mathcal{G}$) makes R weakly coarser, while the elimination of data (that is, the reduction of $\mathcal{G}$) makes R weakly finer. Thus, to usefully refine the welfare relations, one must either eliminate or reinterpret data. Accordingly, if there is one GCS in which x is chosen over y, and another in which y is chosen over x, we look for objective criteria that might allow us to officiate between these GCSs, with the object of discarding or reinterpreting one of them. We can then construct a new welfare relation, R', based on the revised choice correspondence, which may be finer than R and contain fewer welfare optima. The same comments apply for P^*.

How might we officiate between conflicting GCSs? One seemingly natural possibility, which we call "self-officiation," is to officiate based on choices. If the individual makes conflicting choices for two GCSs, $G_1 = (X, d_1)$ and $G_2 = (X, d_2)$, simply allow him to choose between these GCSs. This, however, creates another GCS, call it $G_3 = (X, d_3)$. Since the expansion of $\mathcal{G}$ to include G_3 does not refine usefully either the welfare relation or the sets of welfare optima, we have an "irresolvability principle"—new choices cannot resolve normative ambiguities associated with existing choice conflicts. To officiate, we must therefore rely on nonchoice data.

B. *Refinements Based on Information Processing*

When we say that an individual's standard choice situation is X, we mean that, based on all of the objective information that is available to him, he is actually choosing among elements of X. In standard economics, we use this objective information to reconstruct X, and then infer that he prefers his chosen element to all the unchosen elements of X. But what if he fails to use all of the information available to him, or uses it incorrectly? What if the objective information available to him implies that he is actually choosing from the set X, while in fact he believes he is choosing from some other set, Y? In that case, should a planner nevertheless mimic his choice when making a selection from X? Not in our view.

Why would the individual believe himself to be choosing from some set, Y, when in fact, according to the available objective information, he is choosing from the set X? 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. He may have learned from his past experiences more slowly than the objective information would permit. Thus, the operations of these cognitive processes pertain to the question of whether, at the moment of choice, he appreciates that he is choosing from X.

In principle, if we understand the individual's cognitive processes sufficiently well, we may be able to identify his perceived choice set Y, and to reinterpret the choice as pertaining to the set Y rather than to the set X. We refer to this process as "deconstructing choices." While it may be possible to accomplish this in some instances (see, e.g., Botond Kőszegi and Matthew Rabin 2007), we suspect that, in most cases, this is beyond the current capabilities of science. We nevertheless submit that there are circumstances in which nonchoice evidence can reliably establish the existence of a significant *discrepancy* between the actual choice set, X, and the perceived choice set, Y. This occurs, for example, in circumstances where it is known that attention wanders, memory fails, forecasting is naive, and/or learning is slow. In these instances, we say that the GCS is suspect.

We propose using nonchoice evidence to officiate between conflicting choice data by deleting suspect GCSs. Thus, for example, if someone chooses x from X under condition d', where he is likely to be distracted, and chooses y from X under condition d'', where he is likely to be focused, we would delete the data associated with (X, d') before constructing the welfare relations. In effect, we take the position that (X, d'') is a better guide for the planner than (X, d'). Even with the deletion of choice data, the welfare relations may remain ambiguous in many cases due to other unresolved choice conflicts, but they nevertheless become (weakly) finer, and the sets of welfare optima grow (weakly) smaller.

What types of nonchoice evidence might one use to determine the circumstances in which internal information processing systems work well, and the circumstances in which they work poorly? Evidence from neuroscience concerning the functioning of various cognitive processes can potentially shed light on the operation of processes governing attention, memory, forecasting, and learning. This evidence can provide an objective basis for determining whether a particular choice situation is suspect. For example, if memory is shown to function poorly under certain environmental conditions, GSCs that are associated with those conditions, and that require factual recall, are suspect. Our work on addiction (Bernheim and Rangel 2004) provides a more elaborate illustration. Citing evidence from neuroscience, we argue that the habitual use of addictive substances causes specific information processing systems to malfunction under identifiable circumstances. The choices made in these circumstances are therefore suspect, and welfare evaluations should be guided by choices made in other circumstances. More generally, these observations define a *normative* agenda for the emerging field of neuroeconomics.

In many situations, simpler forms of evidence may suffice. If, for example, an individual characterizes a choice as a mistake on the grounds that he neglected or misunderstood information, this may provide a compelling basis for declaring the choice suspect. Other considerations, such as the complexity of a GCS, could also come into play.

REFERENCES

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

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

Bernheim, B. Douglas, and Antonio Rangel. 2007. "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." Unpublished.

Caplin, Andrew, and John Leahy. 2001. "Psychological Expected Utility Theory and Anticipatory Feelings." *Quarterly Journal of Economics*, 116(1): 55–79.

Gul, Faruk, and Wolfgang Pesendorfer. 2001. "Temptation and Self-Control." *Econometrica*, 69(6): 1403–35.

Kőszegi, Botond, and Matthew Rabin. 2007. "Revealed Mistakes and Revealed Preferences." Unpublished.