CHAPTER 4

ON TWO POINTS OF VIEW REGARDING REVEALED PREFERENCE AND BEHAVIORAL ECONOMICS

RAN SPIEGLER

The concept of revealed preferences surfaces quite often in methodological debates over behavioral economics. There seems to be a close linkage between economists' attitudes to the revealed preference principle and the opinions that they hold regarding paternalistic policies, or their relative evaluation of decision models. This chapter is an attempt to clarify several aspects of this linkage.

The revealed preference principle is part of a philosophical tradition that places restrictions on professional discourse, by judging parts of it to be "meaningless." It was originally formulated at a time when economic models involved little more than static choice of a consumption bundle from a budget set. In this simple consumer-theoretic environment, a single "mental construct" is attributed to the decision maker, namely, a utility function $u : X \to R$. The revealed preference principle then means that any property of u is meaningful only if it can be defined in terms of a preference relation $R \subseteq X \times X$, which in turn can be elicited from choice experiments whose outcomes are summarized by a choice correspondence c, where $c(A) \subseteq A$ for every nonempty $A \subseteq X$. To quote Gul and Pesendorfer (chapter 1):

> In the standard approach, the terms "utility maximization" and "choice" are synonymous. A utility function is always an ordinal index that describes how the

> individual ranks various outcomes and how he behaves (chooses) given his constraints (available options). The relevant data are revealed preference data, that is, consumption choices given the individual's constraints.

To illustrate the principle in action, consider an economist who discusses a policy issue and employs the "mental construct" of utility for this purpose. For whatever reason, the economist's policy analysis hinges on whether an agent's utility function is concave or convex. The revealed preference principle implies that the economist's distinction is meaningful only if concave and convex utility functions are distinguishable on the basis of the agent's observable choices in the domain of choice problems that are relevant for the policy issue. If they are not, then an adherent of the revealed preference principle would dismiss the economist's conclusions from the policy analysis.

As decision models became more complex, additional "mental constructs" entered economic analysis. For instance, subjective expected utility theory involves *two* mental constructs: utility rankings and subjective probability. It has been pointed out [see Karni, 2005] that in subjective expected utility theory, subjective probability and state-dependent utility cannot be distinguished behaviorally. Thus, even when we remain safely within the boundaries of what Gul and Pesendorfer call "mindless economics," the "revealed preference justification" for the concept of utility is more nuanced than in the basic consumer-theoretic environment.

When we reach the "nonstandard" decision models that appear in the bounded rationality and behavioral economics literatures, the tension with the revealed preference principle becomes much stronger. Here are a few examples of such models:

- A choice procedure that selects the median alternative from each choice set, according to a linear ordering of the grand set of alternatives
- A utility maximization model, in which the sole carrier of utility is the decision maker's subjective belief
- A multi-selves model, in which an extensive form decision problem is modeled as an intrapersonal game, and the decision maker's choice is the outcome of subgame perfect equilibrium in that game

In all three cases, the primitives of the decision model contain a preference relation or a utility function. However, it is not true that " 'utility maximization' and 'choice' are synonymous." In the first example, the "median" procedure (which captures the "compromise effect" discussed by psychologists—see Simonson [1989]) induces a choice correspondence that violates the Weak Axiom of Revealed Preferences. In the second example, a utility ranking between two subjective beliefs cannot be revealed by an observable act of choice. In the third example, the decision maker's choice is the outcome of an algorithm that involves a collection of utility functions, but not

AQ: Due to space constraints we have shortened the running head. Kindly confirm is it.

by way of maximizing any of them. Thus, any economic analysis—especially one that involves welfare judgments—that makes use of the concept of utility inherent in these models is inconsistent with the revealed preference principle in its narrow formulation, which finds expression in the statement that "'utility maximization' and 'choice' are synonymous."

One possible response to this state of affairs is that the concept of revealed preferences was invented when economists lacked tools for direct measurement of mental constructs such as utility. Nowadays we have better tools. If, thanks to these tools, a theorist puts a lot of faith in a nonstandard decision model, she can discard the revealed preference principle altogether. In particular, she should not allow the principle to interfere with a potentially useful welfare analysis, which happens to be based on a concept of utility that appears in the description of the model yet fails to be synonymous with observable choices.

Another possible response is to discredit the nonstandard models. According to this point of view, since utility maximization and observed choice are not synonymous in these models, an economic analysis (especially one that involves welfare judgments) that relies on the concept of utility inherent in these models is inadmissible. Either economists give up the ambition to incorporate the psychological factors that these models purport to capture, or they substitute these decision models with alternative models for which "'utility maximization' and 'choice' are synonymous."

I believe that many economists would associate the first point of view with the behavioral economics approach and the second point of view with the decision-theoretic approach to "psychology and economics"—especially with the "mindless economics" vision portrayed in Gul and Pesendorfer. In this chapter, I address these two viewpoints, and the bigger question that they reflect: What is the role of the revealed preference principle in nonstandard decision models that incorporate novel psychological factors? My approach is to examine more specific, concrete claims, which I perceive as being representative of the two respective viewpoints.

Claim 4.1.

> Behavioral economists can "safely" express decision models in the language of utilities, without trying to characterize them in terms of general properties of their induced choice correspondences. This "revealed preference exercise" can be left entirely for specialized decision theorists.

This claim is implicit in the modeling practice of behavioral economics. Rarely do we see a paper that can be identified as a "behavioral economics" paper, in which a choice-theoretic characterization exercise, however rudimentary, is carried out. This avoidance may be a by-product of a rejection of the status of the revealed preference principle as a philosophical foundation for decision models.

However, the explanation may also be "cultural": in its style and rhetoric, behavioral economics is closer to "applied theory" than to "pure theory" and, as such, perpetuates the traditional division of labor between applied theorists and decision theorists.

Let us turn to the second claim. In a series of articles, Gul and Pesendorfer [2001, 2004, 2005a] present a modeling approach to dynamic decision making in the presence of self-control problems, and contrast it with the multi-selves approach. In particular—and this is most relevant for the present discussion—they argue that the former approach is consistent with the revealed preference principle, whereas the latter approach is not. Gul and Pesendorfer's claim can be summarized as follows:

Claim 4.2.

Multi-selves models of dynamic decision making are inconsistent with the revealed preference principle. Moreover, they can and should be reformulated as (or substituted by) decision models for which utility maximization and choice are synonymous.

This essay is critical of both claims. Contrary to claim 4.1, I argue that a rudimentary revealed preference exercise is a highly useful "diagnostic" tool for economists who develop nonstandard, "behavioral" decision models, independently of their attitude to the principle of revealed preferences as a philosophical foundation for decision models. And contrary to claim 4.2, I argue that the narrow version of the revealed preference principle has limitations as a theory-selection criterion in the development of decision models with a "rich" psychology.

The rest of this chapter is structured as follows: First I present a critical discussion of claim 4.1. In order to make the analysis as concrete as possible, I use the literature on "utility from beliefs," which has received wide attention in recent literature, as a test case. Then I discuss claim 4.2, in the context of Gul and Pesendorfer's model of self-control preferences. Last, I summarize my "lessons" from these analyses.

Revealed Preferences as a "Diagnostic Tool": The Case of "Utility from Beliefs"

When theorists embed a decision model in a larger model of an economic environment, they typically express the decision model in the language of utilities, without bothering to characterize it first in terms of general properties of its induced choice

correspondence. This modeling practice is very efficient. However, it tends to make it difficult to perceive general properties of the behavior implied by the decision model. Partly because of this difficulty, there is a specialized branch of economic theory, namely, decision theory, whose job is to clarify, via "representation theorems," how properties of a utility function are defined in terms of revealed preferences. When an applied theorist writes down a model in the language of utilities, he is typically "reassured" that a decision-theoretic analysis that "permits" him to use the model has already been carried out.

In behavioral economics, the difficulty is intensified for several reasons. First, "behavioral" models sometimes retain the concept of utility yet abandon utility maximization, which makes it harder to grasp the relation between choice behavior and the shape of the utility function, Second, even when utility maximization is retained, the domain over which utility is defined is often unusual and complicated. Finally, the behavioral economist typically writes down the decision model before the "reassuring" decision-theoretic exercise has been conducted.

This certainly does not lead me to conclude that proposing a "behavioral" model must be accompanied by a standard decision-theoretic exercise, which most conspicuously includes axiomatization. However, in this section I argue that a rudimentary revealed preference exercise is *heuristically valuable* for the behavioral economist. The reason is that it may serve as a safeguard against misleading interpretation of the model's assumptions, domain of applicability, and conclusions.

To substantiate this claim, I discuss it in the context of the literature on "utility from beliefs," which has received wide attention recently. The basic idea underlying this literature is that people's well-being is often directly affected by their beliefs. For example, a decision maker may derive direct satisfaction from anticipation of high material payoffs; or, she may suffer a disutility (called "cognitive dissonance") if her belief fails to rationalize her actions. In games, a player's sense of disappointment at his opponent's lack of generosity often depends on how he expected the opponent to behave in the first place.

Models with utility from beliefs provide an ideal test case for claim 4.1. On one hand, the idea that beliefs directly affect our sense of well-being is highly intuitive. This intuition is supported by a rich psychology literature that documents the emotional effects of beliefs and how they sometimes lead decision makers to self-serving distortion of their beliefs. On the other hand, the idea of "utility from beliefs" obviously runs counter to the narrow version of the revealed preference principle, since no choice experiment can directly reveal the utility ranking between two subjective beliefs.

For an adherent of the narrow version of revealed preferences, economic analysis that is based on a utility-from-beliefs model (especially welfare analysis) is inadmissible. But does it follow that if one does not share this dismissive attitude to the concept of utility from beliefs, one can entirely dispense with the revealed preference exercise when analyzing a utility-from-beliefs model? In this section I

describe a number of recently proposed utility-from-beliefs models and show how the notion of revealed preferences has value as a "diagnostic tool," which is independent of one's opinion regarding the principle's status as a philosophical foundation for decision models.

Self-Deception

I begin by examining the model of "optimal expectations" due to Brunnermeier and Parker [2005] (henceforth BP). This is a model of self-deception, according to which subjective belief is a *choice variable*. When people choose how much to distort their beliefs, they trade off the anticipatory gains from holding an overoptimistic belief against the material loss that results from making a decision that is based on an incorrect belief.[1]

Whereas BP analyze a rather complicated model with a long horizon, here I examine a stripped-down, two-period version of their model. Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be a set of states of nature where $n \geq 2$, and let A be a set of feasible actions. Let $u(a, \omega)$ be the material payoff that the decision maker (henceforth DM) derives from action a in state ω. Let $q(\omega)$ be the objective probability of ω. Finally, let $\alpha \in (0, 1)$.

The DM's decision rule is to choose a belief p and an action a to maximize the following expression:

$$\alpha \cdot \sum_{\omega \in \Omega} p(\omega) \cdot u(a, \omega) + (1 - \alpha) \cdot \sum_{\omega \in \Omega} q(\omega) \cdot u(a, \omega) \tag{4.1}$$

where a maximizes $\sum_{\omega \in \Omega} p(\omega) \cdot u(a', \omega)$, and the support of p is contained in the support of q. Let c_{BP} denote the choice correspondence induced by the BP decision rule.

The BP model is inconsistent with the narrow version of the revealed preference principle. Expression 4.1, which is intended to represent the DM's "well-being," is defined over a subset of $A \times \Delta(\Omega)$. However, the DM's act of choosing the subjective belief $p \in \Delta(\Omega)$ is unobservable. Note that we could interpret the BP model as if it describes the following two-stage game, in which a "preacher" manipulates a gullible "student" into holding any belief. The reason the preacher can do that is that the student believes that the preacher is absolutely truthful, whereas in fact the preacher conveys information selectively. The preacher moves first by choosing p, and the student moves second by choosing a. The preacher's objective is to maximize expression 4.1, whereas the student's objective is to maximize $\sum_{\omega \in \Omega} p(\omega) \cdot u(a, \omega)$. In this case, the preacher's choice of p corresponds to an observable, selective transmission of information. Under this interpretation, the model is consistent with revealed preferences, but it ceases to be a model of self-deception.

BP's first application of the model concerns risk attitudes. They examine the DM's choice between *two* actions: a safe action and a risky action. In terms of our specification of the model, let $A = \{a_s, a_r\}$, let $u(a_s, \omega) = 0$ for every $\omega \in \Omega$, and suppose that $\sum_\omega q(\omega)u(a_r, \omega) < 0$, yet $\max_\omega u(a_r, \omega) > 0$. Denote $\omega^* \in \arg\max_\omega u(a_r, \omega)$.

A standard DM who maximizes expected material payoffs without trying to deceive himself would play safe (a_s). In contrast, it is easy to see that there exists $\alpha^* \in (0, 1)$, such that for every $\alpha > \alpha^*$, the DM prefers to believe $p(\omega^*) = 1$ and play a_r. Under this belief–action pair, expression (4.1) is reduced to $\alpha \cdot 1 \cdot u(a_r, \omega^*) + (1 - \alpha) \cdot \sum_\omega q(\omega)u(a_r, \omega)$, which is strictly positive if α is sufficiently close to one.

BP conclude that their model implies excessive risk taking. Indeed, some features of this effect appear attractive. People often like to gamble when the maximal loss is very low but the maximal gain is very high, and the explanation may well be that they enjoy the pure anticipation of a large gain. However, recall that BP's conclusion relies on choice problems that involve two actions. What happens when we add a third action to the choice set?

Proposition 4.1.

Fix q and α. Let a_s be a safe action; that is, $u(a_s, \omega) = 0$ for every $\omega \in \Omega$. Then, there exist a material payoff function u and a pair of actions a_r, a'_r, such that $a_r \in c_{BP}\{a_s, a_r\}$ and $c_{BP}\{a_s, a_r, a'_r\} = \{a_s\}$.

Proof

Construct the following payoff function u:

action/state	ω_1	ω_2	$\cdots$	ω_n
a_s	0	0	$\cdots$	0
a_r	1	$-k$	$\cdots$	$-k$
a'_r	m	$-n$	$\cdots$	$-n$

where $k, m, n > 0$, $m > 1$, and k satisfies

$$\alpha + (1 - \alpha) \cdot [q(\omega_1) - k(1 - q(\omega_1)] = 0.$$

Under this specification, $c_{\mathrm{BP}}\{a_s, a_r\} = \{a_s, a_r\}$. To see why, note that when the DM chooses to believe $p(\omega_1) = 1$, the only action in the choice set $\{a_s, a_r\}$ that is compatible with this belief is a_r. But under this belief–action pair, the DM's payoff, given by expression 4.1, attains a value of 0. Clearly, every other belief that is compatible with a_r yields a payoff below 0. Therefore, the only belief that justifies playing a_r is $p(\omega_1) = 1$.

Now suppose that the choice set is $\{a_s, a_r, a'_r\}$. If the DM chooses to believe $p(\omega_1) = 1$, then since $m > 1$, the only action that is compatible with this belief is

a'_r. The DM's payoff under this belief–action pair is

$$\alpha m + (1 - \alpha) \cdot [mq(\omega_1) - n(1 - q(\omega_1)],$$

which is strictly negative if n is sufficiently large. For every other belief that is compatible with a'_r, the DM's payoff will be even lower. Thus, as long as n is sufficiently large, the DM necessarily chooses the action a_s together with any belief that is compatible with it (e.g., $p = q$). ■

Thus, the choice correspondence implied by the BP model violates Independence of Irrelevant Alternatives (IIA). For every objective probability distribution over Ω and for every α, one can construct a material payoff function such that the DM's tendency to choose a (materially inferior) risky action can be undone by adding another risky action to the choice set.[2]

The intuition for this IIA violation is interesting. In the two-alternative choice problem, when the DM chooses the risky action a_r, she might as well deceive herself into thinking that $p(\omega^*) = 1$. This huge overoptimism justifies taking the risky action, because the material decision loss is outweighed by the large anticipatory gain. However, when the risky action a'_r is added to the feasible set, excessive optimism would cause the DM to choose a'_r. Given q and u, the expected material loss from a'_r is so big that it outweighs the anticipatory gain. Thus, in order to induce a_r, the DM must restrain her overoptimism, but this means that the anticipatory gain is not high enough to justify this action anymore. Therefore, the DM ends up choosing the safe action.

I find this quite insightful and certainly worthy of further study. It may well turn out to shed light on how decision makers' propensity for self-deception varies with their set of options. At any rate, the result greatly qualifies BP's claims regarding the risk attitudes implied by their model. The finding that the DM's risk attitudes vary with the choice set is clearly more fundamental than the finding that the DM displays excessive risk seeking in some choice problems with a safe asset and a risky asset.

Had BP approached the issue a bit more like choice theorists, they would have attached greater importance to checking whether c_{BP} satisfies IIA:

- If IIA is satisfied, then the model is behaviorally indistinguishable from a standard model in which the DM maximizes a utility function over actions. This raises an interesting question: Suppose that two decision models account for the same behavior in a prespecified domain of choice problems, yet only one of them is consistent with the narrow version of the revealed preference principle (in the sense that utility maximization is synonymous with observed choice). How should we respond to such behavioral equivalence?

- If IIA is violated, then any statement about the DM's choice under uncertainty has to be qualified because adding an "irrelevant alternative" may reverse the DM's choices between risky and safe actions.

In this case, it appears that a rudimentary revealed preference exercise is valuable, even if the ultimate goal is to develop an "applied" model. I do not think that this little exercise should be left for some future decision theorist who may take it upon himself to axiomatize the BP model. Instead, it should accompany the original development of the model. The question of whether the choice correspondence induced by a decision model satisfies IIA is so basic that readers of an article that presents the model for the first time would probably like to know about it.

This example illustrates that thinking in terms of revealed preferences is "diagnostically" valuable for the development of a "behavioral" decision model. In particular, reexpressing the model in the language of choice correspondences brings to the fore key behavioral properties, which are known to be insightfully linked to properties of utility, and yet are sometimes obscured by the utility language.

Information Acquisition

The primitives of a "revealed preference" model are choice correspondences or preference relations, and the modeler's problem is to relate properties of these objects to properties of the utility representation she is interested in. This encourages the researcher to search for choice problems that seem most likely a priori to elicit the psychological factors under study. For instance, choices between insurance policies are intuitively a good place to look for elicitation of risk aversion; choices between menus are a good place to look for elicitation of a DM's awareness of his self-control problems. Thus, thinking in terms of revealed preferences about a decision model involves looking for domains of choice problems that are promising in the sense that they intuitively seem capable of eliciting the psychology captured by the model.

The BP utility-from-beliefs model is meant to be about self-deception—namely, about the way people manipulate their own beliefs. Although the model assumes that the DM directly chooses what to believe, people can also distort their beliefs indirectly, through their choice of information sources. Indeed, introspection and casual observation suggest that the phenomenon of self-deception has a lot to do with aversion to potentially unpleasant information. People who try to deceive themselves are also likely to hire "yes men" as advisors, flip TV channels in order to avoid disturbing news, and keep "dangerous" books out of their home. In particular, there is a strong intuition that people's choice of a biased information source over another (e.g., watching Fox News rather than BBC News) is often indicative of the kind of self-serving distortion of beliefs they

are trying to attain. Finally, this way of influencing one's beliefs is observable in principle.

Thus, according to the prescription given at the top of this subsection, we should be interested in examining how a DM who derives direct utility from anticipated payoffs chooses information sources. BP leave information acquisition outside their model. In order to extend the model to this domain, we need to add a preliminary stage to the decision process, in which the DM chooses a signal from a set of feasible signals S. Given the realization of the chosen signal, the DM chooses a belief and an action as in the original BP model, so as to maximize expression 4.1.

This description is incomplete, because it does not tell us how the DM updates her beliefs, or how the updating process interacts with the DM's direct choice of beliefs. I cannot think of any assumption that would not sound artificial in this context. However, the most standard assumption would be that the DM updates her beliefs according to Bayes's rule upon the realization of a signal, so that the subjective belief p is chosen given the updated objective probability distribution.

Proposition 4.2.

Fix A, and suppose that S consists of signals that never rule out any state with certainty. Then, the DM's behavior throughout the two-stage decision problem—and particularly his choices between signals in the first stage—is indistinguishable from those of a standard DM who tries to maximize the expectation of some utility function $v(a, \omega)$.[3]

Proof

Given a, let P_a be the set of subjective beliefs p for which

$$a \in \arg\max_{a' \in A} \sum_{\omega \in \Omega} p(\omega) u(a, \omega).$$

Let $p_a^* = \arg\max_{p \in P_a} \sum_{\omega \in \Omega} p(\omega) u(a, \omega)$. Define

$$v(a, \omega) = (1 - \alpha) \cdot u(a, \omega) + \alpha \cdot \sum_{\omega \in \Omega} p_a^*(\omega) u(a, \omega).$$

Note that the second term in this expression is a function of a only. Thus, in the second stage of the two-stage decision problem, a DM who chooses p and a according to the BP decision rule behaves as if she chooses a to maximize the expectation of v.

Let us turn to the first stage. Recall that we restrict attention to signals that never rule out any state with certainty. Therefore, the set of feasible subjective beliefs p in the second stage remains the set of probability distributions over Ω whose support

is weakly contained in the support of q. Therefore, the DM chooses signals in the first stage as if his objective is to maximize the indirect utility function induced by the maximization of the expectation of v in the second stage. ■

This result means that under a slight restriction of the domain of feasible signals, the DM is never averse to information.[4] But if the most standard extension of the BP model cannot capture preference for biased information sources, which intuitively seem to originate from a desire to attain self-serving beliefs, what does it mean for the self-deception interpretation of the BP model? In particular, how should we regard the theory that a financial investor's excessive risk taking is due to self-deception, when we know that according to the same theory, she never rejects potentially unpleasant information?

Note that under the "indoctrination" interpretation suggested in the preceding subsection, there is nothing strange about the observation that the preacher is never averse to information. Of course, the economic situations that fit the "indoctrination" interpretation are quite different from those that fit the "self-deception" interpretation.

As in the preceding subsection, a "revealed preference approach" has value as a diagnostic tool. Specifically, looking at choices of signals—a domain of choice problems that seems capable of exposing the psychology of self-deception—may lead us to question the original interpretation of the model and its domain of applicability.

Other Utility-from-Beliefs Models

In the extended BP model of the preceding subsection, the DM distorts his beliefs both through self-deception and through choice of signals. In another class of utility-from-beliefs model [e.g., Caplin and Leahy, 2004; Köszegi, 2003, 2006], the DM's belief enters as an argument in his utility function, but he does not get to choose what to believe. The only way he can affect his beliefs is through information.

The decision process that is embedded in these models consists of two stages. In the first stage, the DM chooses a signal and updates her beliefs according to the signal's realization, via Bayes's rule. In the second stage, the DM chooses an action a, given her updated belief. A common application of the model involves a patient who faces a choice between diagnostic tests having different probabilities of a false positive and a false negative. Another application may involve a media consumer who chooses a TV news channel in order to be informed about the latest Middle East crisis.

In these models, the DM's objective function typically takes the following form:

$$M(p) + \sum_{\omega \in \Omega} p(\omega) \cdot u(a, \omega), \tag{4.2}$$

where u is a material payoff function and $p \in \Delta(\Omega)$ is the DM's posterior belief, arrived at from the prior q and the observed signal, via Bayesian updating. The second term in the DM's objective function represents his expected material payoff. The term $M(p)$ is a continuous function, which is meant to represent the "anticipatory payoff" associated with a posterior belief p (details of u may enter the specification of M). Depending on the shape of M, the DM may display aversion to information in the first stage.

The DM's indirect expected utility conditional on the posterior belief p can be written as follows:

$$U(p) = M(p) + \max_{a \in A} \sum_{\omega \in \Omega} p(\omega) \cdot u(a, \omega)$$

Thus, as long as we are only interested in the DM's first-stage behavior, we may adopt a reduced-form model, according to which the DM chooses a signal that maximizes the expectation of $U(p)$, given her prior q, knowing that she will update q according to Bayes's rule.[5]

Eliaz and Spiegler [2006] carry out a rudimentary revealed preference analysis of this reduced form model. Note that the DM's first-stage choices between signals are *indexed* by the parameter q. If the DM's choices of signals are rational, the first-stage "choice data" can be represented by a profile of preference relations $(\succsim_q)_{q \in \Delta(\Omega)}$ over the set of signals S. The problem is to account for the choice data by the reduced-form model.

Eliaz and Spiegler [2006] point out several difficulties with this model. First, if U accounts for the choice data, then so does the function $V(p) = U(p) - pc^T$, where $c = (c_1, \ldots, c_n)$ is a vector of real numbers. In particular, we can choose c such that there will be two beliefs, p and p', for which $U(p) > U(p')$ and $V(p) < V(p')$. In some special cases, we can choose c such that U and V will induce opposite rankings for *any* pair of beliefs.

Why does this little result pose an interpretational difficulty? In their article on economic implications of cognitive dissonance, Akerlof and Dickens [1982: 307] state: "[P]ersons ... manipulate their own beliefs by selecting sources of information likely to confirm "'desired beliefs.'" Bénabou and Tirole [2002: 906] write in a similar vein: "[P]eople just like to think of themselves as good, able, generous, attractive, and conversely find it painful to contemplate their failures and shortcomings." Indeed, there is an intuition that people's preference for information sources having a particular bias is related to their perception of what constitutes a desired belief.

For instance, we may expect that when a media consumer chooses whether to be informed about the latest Middle East conflict by watching Fox News or BBC News, his decision will be linked to the kind of narrative he wants to believe in (which of the parties to the conflict is to blame, or which party won). The fact that U and

V may represent the same choices between signals implies that two different media consumers, having diametrically opposed views as to what constitutes a desired narrative, could end up watching the same channel. Thus, contrary to the intuition articulated by Akerlof and Dickens [1982], the distinction between "desired" and "undesired" beliefs may be irrelevant to the DM's choice of information sources, according to the above utility-from-beliefs model.

Comment

Caplin and Leahy [2004] make essentially the same observation, but they do not seem to share my view that it is a cause for concern. Instead, they highlight a different implication. Suppose that the DM is the receiver in a communication game, in which the sender is a "concerned expert" whose objective is to maximize the receiver's expected utility from beliefs. A "signal" in this model represents the sender's information transmission strategy. In this environment, U and V may be distinguished behaviorally, because they imply different disclosure incentives for the sender, and therefore may induce different equilibrium outcomes. I agree with this claim, but find it independent of the above interpretational difficulty.

Another difficulty pointed out by Eliaz and Spiegler [2006] is that the model fails to account for a variety of realistic prior-dependent attitudes to information, which intuitively seem to originate from anticipatory feelings. For instance, suppose that the DM ranks the fully informative signal above all other signals when $q(\omega_1)$ is close to 1. Then, according to the model, this ranking must hold for all priors. It follows that the model cannot capture the behavior of a patient who wants to have full knowledge of her medical condition when she is quite sure that she is in good health, yet does not want to know the whole truth when she is not so sure.

Both difficulties raise doubts regarding the model's ability to explain anomalous attitudes to information, despite the intuition that such attitudes sometimes originate from the direct effect of beliefs on the DM's well-being. It should be emphasized that these difficulties can be discovered without abandoning the "applied theory" manner of expressing models purely in the language of utilities. Nevertheless, a rudimentary revealed preference exercise makes it easier to notice them. One ingredient of this exercise is to check whether the utility representation is unique with respect to certain transformations. This makes the first difficulty easy to discover. In addition, once the "choice data" are written down systematically, it becomes obvious that the DM's choices over signals are indexed by the prior q, whereas the%enlargethispage11pt utility function U is not indexed by q. This observation makes it more urgent for the theorist to seek connections between the DM's choices of signals at different priors, which makes the second difficulty easy to discover.

Summary

The point of this section is simple. Even if one rejects the revealed preference principle as a criterion for determining the admissibility of "behavioral" decision models, the principle still has value in the development of such models. A rudimentary revealed preference exercise helps clarifying general aspects of the behavior induced by the model. The clarification obtained in this way is so basic that it cannot be left for a future decision theorist. Instead, it should be part of the behavioral theorist's bag of tools.

Rubinstein and Salant (chapter 5) make a related comment. They argue that although certain choice procedures cannot be described as the outcome of utility maximization, they can be characterized and differentiated from one another by properties of their induced choice correspondence. Thus, a revealed preference exercise is instrumental in characterizing "nonstandard" decision models that incorporate novel psychological factors.

Revealed Preferences as a Theory-Selection Criterion: The Case of Self-Control Preferences

In the preceding section I argue in favor of a modeling approach that borrows the "diagnostic" aspect of the revealed preference principle. Claim 4.2 described in the introduction enunciates a more ambitious view of the role of the revealed preference principle in behavioral economics. According to this claim, if a decision model fails to satisfy the property that "'utility maximization' and 'choice' are synonymous," then it is an inferior model, in the sense that it cannot provide a basis for welfare analysis that is consistent with the revealed preference principle. Therefore, the model should be reformulated as, or substituted by, a model for which utility maximization and choice are synonymous.

This position has been articulated most insistently by Faruk Gul and Wolfgang Pesendorfer (GP henceforth) in a series of papers. For instance, in GP [2005b], they propose a theory of social preferences, in which a player's preferences over strategy profiles in a game depend on his and his opponent's types, where a player's type represents his "personality." GP pit this theory against models of social preferences based on the formalism of psychological games (due to Geanakopolos, Pearce, and Stachetti [1989]), in which players' payoffs are also a function of their hierarchy of beliefs regarding the strategy profile. The two theories are meant to cover the same terrain (social preferences), yet one is a "revealed preference theory," whereas the other is not.

Such comparisons have received the widest attention in the context of GP's highly original decision-theoretic modeling approach to changing tastes and self-control (which builds on foundations laid by Kreps [1979], Dekel, Lipman, and Rustichini [2001], and Kreps and Porteus [1978]). GP compare this approach to the multi-selves approach. In GP [2005a: 430], they argue: "The advantage of our approach is that preferences over decision problems are—at least in principle—observable. Rather than speculate about the appropriate model of expectation formation, we offer choice experiments that identify Strotz's model of behavior." Epstein [2006: 4] endorses this view:

> Gul and Pesendorfer [2005a] describe advantages of their approach. One is that the axiomatic method permits identification of the exhaustive empirical content of the model, expressed through restrictions on in principle observable behavior at a fixed time 0. In contrast, Strotz's multi-selves interpretation involves hypotheses not only about time 0 behavior, but also about expectations of future behavior.

Thus, GP's position seems to be as follows: The decision-theoretic modeling approach is consistent with the revealed preference principle, in the sense that all relevant "mental constructs" (not only utility rankings, but also the DM's expectations of future behavior) are revealed by observable choices. In contrast, the multi-selves approach is inconsistent with the revealed preference principle, in the sense that it relies on a priori assumptions regarding the DM's formation of expectations of future behavior. In this section I discuss the GP modeling approach in light of these claims. In particular, I ask whether it jibes with the revealed preference principle in a way that the multi-selves approach fails to.

In the simplest form of the GP model of self-control preferences [GP 2001], the DM goes through a two-period decision process. In the first period, she chooses a menu of lotteries. In the second period, she selects a lottery from the chosen menu. GP assume that the DM has complete, transitive preferences over menus, which can be represented by the following utility function:

$$U(A) = \max_{x \in A}[u(x) + v(x)] - \max_{y \in A} v(y),$$

where u and v are expected utility functions. As far as first-period behavior is concerned, the GP model definitely meets the requirement that " 'utility maximization' and 'choices' are synonymous."

But the interest in the GP model lies precisely in the interpretation of the menu as a choice set, and of the function U as an indirect utility. In particular, one of the components in the utility representation, namely, $\arg\max_{x \in A}[u(x) + v(x)]$, is interpreted as the lottery that the DM expects to choose in the second period. Moreover, when applying this model, GP and other practitioners assume that these expectations are correct. This interpretation is crucial for the applications of the

model. Yet, it is *external* to the revealed preference exercise. We could just as well assume that the DM's expectations are systematically incorrect, without changing anything in the analysis of "time 0" preferences over menus.

GP [2005a] appear to be saying the same thing, in reference to their "revealed preference theory of changing tastes" (which, in its two-period version, is essentially an axiomatization of the first-period behavior implied by the two-period Strotzian model):

> As in standard models of dynamic choice we view the decision maker as expressing a preference at one point in time (period 0). The representation of these preferences suggests behavior in future periods that can be interpreted as the agent's implicit expectations. Whether these expectations are correct or not (that is, whether the agent is sophisticated or not) can be treated as a separate question. That is, the representation is a valid description of period 0 behavior whether or not the agent has correct expectations, as long as the axioms are satisfied. [432]

The same words could apply to the GP model of self-control preferences. Yet, in light of this statement, the decision-theoretic and multi-selves approaches appear to have the same justification for their assumptions on the DM's first-period expectations of future behavior. Of course, there is a methodological difference between the two modeling practices. The multi-selves approach views as primitive the two selves' preferences over terminal histories of the two-period decision problem (i.e., decision paths) and make explicit assumptions regarding the solution concept that is applied to the extensive decision problem. One can then derive from these assumptions an induced first-period preference relation over menus. The GP modeling practice takes this preference relation as primitive.

But this simply means that the two approaches complement each other. In particular, in line with my claim above, GP's revealed-preference exercise enables a deeper understanding of the decision model. This is quite distinct from claiming that one modeling approach is philosophically more appealing than another because it is consistent with the revealed preference principle in a way that the other approach fails to be. The revealed preference principle cannot be used as a criterion for selecting between the two modeling approaches.[6]

Preferences over Decision Paths

GP [2001] propose an alternative approach to modeling two-period dynamic decision making. The DM is now assumed to have stable *extended* preferences over *decision paths*, instead of first-period menus. Let us denote a decision path by a pair (A, x), where A is a menu and $x \in A$ is a lottery. Let $c(A)$ represent the set of lotteries which the DM actually chooses in the second period, conditional on facing the menu A. GP impose axioms on the DM's preferences over decision paths, which imply the same choice behavior as their original formulation.

In standard models, we do not distinguish between two decision paths with the same chosen elements, (B, x) and (A, x), because we assume that the chosen element is all that matters to the DM. However, when self-control issues are relevant, the decision paths $(\{x, y\}, x)$ and $(\{x\}, x)$ are not necessarily equivalent, because choosing x over y may require the DM to exercise self-control, whereas self-control is not called for when x is the only feasible element.

If we insist on the requirement that "'utility maximization' and 'choice' are synonymous," we must define the DM's preferences over decision paths entirely in terms of choice experiments that directly reveal these preferences.[7] Yet, what *is* the choice experiment that directly reveals the ranking $(A, x) \succ (B, y)$ when $x \notin c(A)$—that is, when the DM never chooses x from A out of his own free will? The only choice experiment that can directly reveal such a ranking involves a choice between *committing* to the decision path (A, x) and *committing* to the decision path (B, y). However, if the DM is able to commit to a path, then the interpretation of A and B as choice sets disappears, and self-control considerations become irrelevant. Consequently, a decision model that incorporates self-control issues cannot be revealed by such choices.[8]

We see that the alternative decision-theoretic approach to self-control is inconsistent with the narrow version of the revealed preference principle, in the sense that some of the utility rankings assumed by the model are not synonymous with observable choices. I do not think that this failure is specific to the GP model. Rather, it seems that the very nature of self-control considerations makes it difficult to provide a complete revealed-preference justification for a model of self-control that assumes a utility ranking over decision paths.

Comment: Game-Theoretic Applicability of the Extended GP Model

The question of what it means for the DM to rank (A, x) over (B, y) when $x \notin c(A)$ is especially important if we wish to embed the extended GP model in *game-theoretic* environments. Consider the following scenario: In the beginning, player 1 is on a diet, with only broccoli at her disposal. She is offered a gift consisting of a cash-equivalent voucher as well as a piece of tempting chocolate. She chooses whether to accept this gift. If she does, then she later has to choose whether to eat the chocolate or exercise self-control and have broccoli. Player 2 has the power to veto a decision to eat chocolate. He chooses whether to exercise his veto power, *after* player 1's first action and *before* his second action. However, player 1 does not observe player 2's decision. Figure 4.1 shows the extensive form of this game.

If we wish to apply the extended GP self-control preferences, then the description of a consequence for player 1 is as written under the terminal nodes of the extensive game. Suppose that player 1 expects player 2 not to exercise his veto power. In this case, player 1 thinks that she has real choice between chocolate and broccoli when she accepts the gift. The feasible consequences for her are $(\{b\}, b)$ (in case

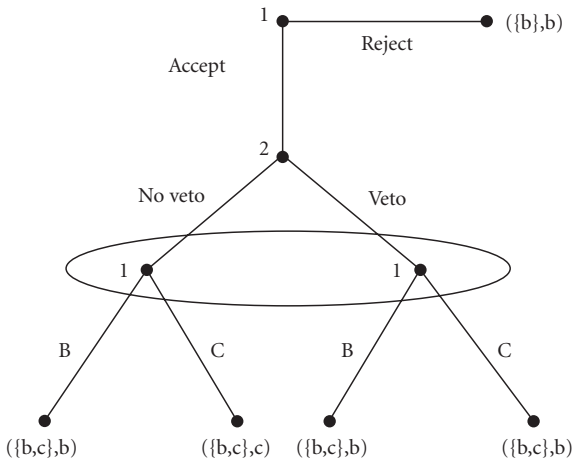

Figure 4.1. A self-control game

she rejects the gift), $(\{b, c\}, b)$ (in case she accepts the gift and eats broccoli), and $(\{b, c,), c)$ (in case she accepts the gift and eats the chocolate). Assume that given her expectations, player 1 chooses to accept the gift and plans to eat the chocolate. If this plan is carried out, it reveals the rankings $(\{b, c\}, c) \succ (\{b\}, b)$ and $(\{b, c\}, c) \succ (\{b, c\}, b)$.

However, suppose that player 2 surprises player 1 by deviating and exercising his veto power. Then, the consequence of this game is now $(\{b, c\}, b)$. The alternative b is not chosen from the set $\{b, c\}$, but follows from player 2's surprise deviation. Is player 1 better or worse off than if she had rejected the gift in the first place? This question is essential to the welfare analysis of this game—and also to the positive analysis, if player 2's objective is to maximize player 1's utility—yet the answer to this question cannot be given by player 1's revealed preferences.

[Alternatively, suppose that player 1 expected player 2 to exercise his veto power in the first place. Would it be appropriate to describe the outcome of accepting the gift and having broccoli as $(\{b, c\}, b)$? Or, would it be more apt in this case to describe it as $(\{b\}, b)$, given that player 1 *perceives* that she has no real freedom to choose chocolate? More generally, in the description of a consequence, is a "menu" the set of alternatives that the player *perceives* as feasible, given his expectation of the opponent's behavior, or is it the set of alternatives that are *truly* feasible given the opponent's actual behavior?]

One could argue that the above description of a consequence does not fit the scenario, because it treats two different experiences of player 1 as if they are equivalent: eating broccoli from the menu $\{b, c\}$ as a result of a personal choice, and eating broccoli from the same menu against personal choice, as a result of player 2's surprise deviation. If we accept this argument, the implication seems to be that the GP model is inapplicable to this game. What does this mean for the applicability of

the GP model, or any other model of self-control for that matter, in game-theoretic contexts? I leave these questions for future inquiry.

Summary

In this section I examined two decision-theoretic approaches to modeling self-control preferences, proposed by GP. One approach assumes utility maximization over decision problems, while another approach assumes utility maximization over decision paths. We first saw that the first approach is no more consistent with the revealed preference principle than the multi-selves approach, as far as the treatment of expectations is concerned. Next, we saw that the second approach cannot be reconciled with the narrow version of the revealed preference principle. This raises doubts regarding the claim that the principle can be used as a criterion for selecting between different decision models that incorporate "nonstandard" psychological motives such as self-control.

Conclusion

"Revealed preferences" are first and foremost a way of getting a systematic, abstract understanding of decision models, most notably utility-maximization models. A revealed preference exercise allows us to realize which aspects of a mental construct that appears in a model are relevant for behavior in a variety of domains of choice problems. Such a systematic understanding is useful as a safeguard against blind spots that working purely with utility functions often creates.

I have argued that this "diagnostic" value of thinking in terms of revealed preferences is especially high in the case of "behavioral" models, and demonstrated this claim with a discussion of several recent utility-from-beliefs models. In my opinion, a philosophical stance that rejects revealed preferences as a justification for decision models does not mean that the behavioral theorist should discard the revealed preference exercise altogether. A rudimentary revealed preference analysis is a very useful member in the theorist's bag of tools as he develops a nonstandard, "behavioral" decision model.

At the same time, I have argued against what I perceive as an attempt to use the narrow version of the revealed preference principle as a theory-selection criterion in the development of decision models with a "rich" psychology. In the case of models of self-control, I showed that it is hard to discriminate between the multi-selves approach and the decision-theoretic approach on this basis. If we wish to continue relying on some notion of revealed preferences as a basis for welfare analysis, and at the same time admit novel psychological phenomena such as self-control or self-deception into our models, then we face the challenge of redefining the concept

of revealed preferences. This challenge may originate from the psychological phenomena themselves, rather than from any particular model that purports to capture them, or from the kind of data that the economist can observe.

NOTES

This chapter is based on a talk I gave at the Methodologies of Modern Economics conference that took place at NYU in July 2006. I am grateful to the organizers of this conference. I also thank Eddie Dekel, Kfir Eliaz, Barton Lipman, and Ariel Rubinstein for numerous conversations that made this chapter possible. I also thank Andy Schotter for useful comments.

1. Other models in which decision makers directly choose their beliefs are due to Akerlof and Dickens [1982], Eyster [2002], and Yariv [2002].

2. The proposition assumes that both a_s and a_r are chosen from $\{a_s, a_r\}$, in order to facilitate the proof. One could construct a payoff function u for which $C_{\text{BP}}\{a_s, a_r\} = \{a_r\}$ and $C_{\text{BP}}\{a_s, a_r, a_r'\} = \{a_s\}$, but the proof would be more tedious.

3. There is no contradiction between this result and proposition 4.1, since here A is held fixed.

4. BP seem to acknowledge this: "However, without relaxing the assumptions of expected utility theory and Bayesian updating, agents would not choose that uncertainty be resolved later because agents take their beliefs as given" [1109].

5. Of course, U may be induced by other functional forms than the one given by expression (4.2).

6. There is an important *economic* difference between the GP model of self-control preferences and the typical Strotzian model. The latter assumes that the domain of first-period and second-period preferences can be restricted to the set of chosen alternatives, whereas the former allows unchosen alternatives to affect preference rankings. This is indeed one of the important contributions of the GP model. However, it does not engender a fundamental methodological distinction between the two approaches.

7. Some of these rankings can be deduced by using the transitive closure of other, directly observed rankings. However, such deduced rankings are based on introspection rather than observable choices and therefore cannot be viewed as revealed preferences.

8. GP [2001: 1415] partially acknowledge this incompleteness of revealed preferences in this model: "For example, if (A, x) is strictly preferred to (A, y) and (A, z), there is no experiment that can determine the agent's ranking of (A, y) and (A, z)."

REFERENCES

Akerlof, George, and William Dickens. 1982. The Economic Consequences of Cognitive Dissonance. *American Economic Review* 72: 307–319.

Bénabou, Roland, and Jean Tirole. 2002. Self-Confidence and Personal Motivation. *Quarterly Journal of Economics* 470: 871–915.

Brunnermeier, Markus, and Jonathan Parker. 2005. Optimal Expectations. *American Economic Review* 95: 1092–1118.
Caplin, Andrew, and John Leahy. 2004. The Supply of Information by a Concerned Expert. *Economic Journal* 114: 487–505.
Dekel, Eddie, Barton Lipman, and Aldo Rustichini. 2001. A Unique Subjective State Space for Unforeseen Contingencies. *Econometrica* 69: 891–934.
Eliaz, Kfir, and Ran Spiegler. 2006. Can Anticipatory Feelings Explain Anomalous Choices of Information Sources? *Games and Economic Behavior* 56: 87–104.
Epstein, Larry. 2006. Living with Risk. Mimeo, University of Rochester.
Eyster, Erik. 2002. Rationalizing the Past. Mimeo, Nuffield College.
Geanakopolos, John, David Pearce, and Ennio Stachetti. 1989. Psychological Games and Sequential Rationality. *Games and Economic Behavior* 1: 60–79.
Gul, Faruk, and Wolfgang Pesendorfer. 2001. Temptation and Self-Control. *Econometrica* 69: 1403–1436.
———. 2004. Self-Control and the Theory of Consumption. *Econometrica* 72: 119–158.
———. 2005a. The Revealed Preference Theory of Changing Tastes. *Review of Economic Studies* 72: 429–448.
———. 2005b. The Canonical Type Space for Interdependent Preferences. Mimeo, Princeton University.
Karni, Edi. 2005. State-Dependent Preferences. In *The New Palgrave: A Dictionary of Economic Theory and Doctrine,* ed. John Eatwell, Murray Milgate, and Peter Newman. London: Macmillan.
Köszegi, Botond. 2003. Health Anxiety and Patient Behavior. *Journal of Health Economics* 22: 1073–1084.
———. 2006. Emotional Agency. *Quarterly Journal of Economics* 121: 121–156.
Kreps, David. 1979. A Representation Theorem for "Preference for Flexibility". *Econometrica* 47: 565–576.
Kreps, David, and Evan Porteus. 1978. Temporal Resolution of Uncertainty and Dynamic Choice Theory. *Econometrica* 46: 185–200.
Simonson, Itamar. 1989. Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research* 16: 158–174.
Yariv, Leeat. 2002. I'll See It When I Believe It—A Simple Model of Cognitive Consistency. Mimeo, Yale University.