Rational Preference and Evaluation

Daniel M. Hausman

> A strong feeling, directly excited by an object, is felt (except when contradicted by the feelings of other people) as its own sufficient justification—no more requiring the support of a reason than the fact that ginger is hot in the mouth; and it almost requires a philosopher to recognize the need of a reason for his feelings, unless he has been under the practical necessity of justifying them to persons by whom they are not shared. (Mill 1884, 296)

IT IS FITTING TO BEGIN WITH A QUOTATION FROM MILL, a rational choice theorist whose work so eloquently bridges the divide between the humanities and the social sciences. In this passage from his *Examination of Sir William Hamilton's Philosophy*, Mill maintains that it requires no subtle philosophical analysis to recognize that feelings require justification or reasons. Yet contemporary economics and rational choice theory apparently take preferences to be "givens" for which no justification is required or perhaps even possible. Even Jon Elster, who has written so insightfully concerning the perversions of reason, maintains at the beginning of his classic monograph, *Sour Grapes*, "Here desires are the unmoved movers, reflecting Hume's dictum that 'reason is, and ought only to be the slave of the passions" (1983, 4).

Mill is clearly right: desires are subject to rational appraisal, and it takes no fancy philosophy to recognize this. Indeed Hume himself argues that there are "general principles of approbation or blame, whose influence a careful eye may trace in all operations of the mind" (1963, 238). Just as

Daniel M. Hausman is the Herbert A. Simon Professor of Philosophy at the University of Wisconsin–Madison. He was educated at Harvard University and New York University, and received his PhD from Columbia University in 1978. His research has centered on epistemological, metaphysical, and ethical issues lying at the boundaries between economics and philosophy. He is the co-founder and former co-editor of the journal *Economics and Philosophy*. His most important books are *The Inexact and Separate Science of Economics* (1992) and *Economic Analysis, Moral Philosophy, and Public Policy* (2006, with Michael McPherson). He is currently writing a book on preferences.

A Hausman, Daniel. "Rational Preference and Evaluation." *Occasion: Interdisciplinary Studies in the Humanities* 1, no. 1 (October 15, 2009), http://occasion.stanford.edu/node/21.

there are normative standards governing the relationship between belief and evidence, so there are normative standards governing the relations between desires and evidence. My task here will not be to defend this claim, which needs no defense, but to clarify what is involved in the appraisal of preferences and to show that appraising preferences is fully compatible with the basic framework of rational choice theory, though not with some formal presentations of the theory.

In place of a caricatured vision of rational choice theory as indifferent to the complexities of the evaluations of people, circumstances, objects and choices, I shall present a view of rational choice theory as containing important fragments of a general theory of evaluation. Unlike many critics and defenders of rational choice theory, who depict it as entirely mute concerning the rationality of individual preferences and the processes by which they are formed, I shall argue that rational choice theory necessarily contains an account of preference formation. So rational choice theory is not indifferent to the questions of value that have been so central to the humanities, and fruitful dialogue should be possible between rational choice theorists and those interested in the humanities.

The processes whereby humans evaluate things are complicated and obscure, and are probably better understood by artists, musicians, poets, novelists and those who stumbled upon some way to live well than by philosophers. The paper will unfortunately have little to say about the most fundamental evaluations, which are the most mysterious and the most philosophically and humanly interesting. In demonstrating the role of evaluation within rational choice theory, I shall instead focus on the prosaic but crucial task of deriving evaluations of the alternatives among which people choose from evaluations of more general outcomes and states of affairs. Although prosaic, this task is not simple. As the pragmatists, including especially John Dewey, emphasized (1929, chap. 10), evaluative reflection is not unidirectional: In thinking about the means to achieve our ends, we may revise the value we attach to ends rather than relying on those values to rank the immediate objects of choice. I shall argue that, on pain of vacuity, rational choice theorists must commit themselves to accounts of the relations between proximate and distant values.

To make this case, I shall begin in section 1 with some words about everyday explanations of actions before turning in section 2 to formal theories of rational choice, which are, in my view, extensions and elaborations of everyday accounts. I shall argue in sections 3 and 4 that economists and decision theorists are in fact committed to views of preference formation, which leaves room for the rational appraisal of preferences. Section 5 argues that, despite some of his rhetoric, Hume believes that preferences can be evaluated. Section 6 is devoted to a technical objection which is answered in section 7 with an account of how formalizations of rational choice theory can permit rational scrutiny of preferences.

1. RATIONAL CHOICE THEORY AND FOLK PSYCHOLOGY

Some people identify rational choice theory with the views of a few right-wing rational choice theorists. Some of those participating in the Stanford Conference on Rational Choice and the Humanities took rational choice theory to hold that capitalism is the best possible economic system, that the free market is always more efficient that government control, or that individuals single-mindedly attempt to maximize their own gain. Although there are rational choice models that are flattering to capitalism and to free markets and rational choice models in which individuals are depicted as concerned with nothing other than their own gain and relentless in its pursuit, there are also rational choice models which explore the pathologies of capitalism and the

inefficiencies of uncoordinated individual action, and rational choice models in which individuals have a wide range of objectives, self-interested, altruistic, and moral.

Rational choice theory consists of a family of models that reformulate the traditional folk-psychological view that human actions can be explained and predicted by the beliefs and desires of agents. "Desire" in the context of folk psychology is a catch-all including a diverse array of motivating factors—emotions of all sorts, aversions, appetites, feelings of obligations—basically any mental state that "pushes" an agent. Donald Davidson (1963, 685–86) speaks of "pro-attitudes." So when one cold Friday night, a hungry student named Ellen takes a frozen pizza out of a refrigerator, unwraps it, puts it in a stove, and turns knobs on the stove, folk psychologists—including Ellen's friends and Ellen herself—explain Ellen's action by Ellen's beliefs, such as her beliefs that turning the knobs causes the stove to heat the pizza, and by her desire to eat hot pizza.

This sort of explanation is familiar, but not very satisfactory. Ellen might also like her pizza still frozen and crunchy, or she might also have a desire to reheat some left-over meatloaf. Or she might rather skip dinner and keep studying decision theory. What explains her action is not merely desiring to eat hot pizza (plus possessing the requisite beliefs), but desiring to do this as much or more than she wants to do any of the feasible alternatives.

Rational choice theory tightens up the folk psychological account of action by replacing the non-comparative catch-all notion of a "desire" with a comparative catch-all notion of "preference." Rational choice theorists can then explain the little interaction between Ellen and the stove in terms of physical constraints, Ellen's beliefs about the outcomes of the alternative actions she can undertake that Friday night, and her catch-all ranking of those outcomes. A rational choice explanation of the pizza warming shows that Ellen ranks its expected outcome at least as highly as any feasible alternative.

Orthodox variants of rational choice theory in economics and decision theory take an agent's preferences to consist of his or her overall evaluation of some set of alternatives. This set of alternatives need not be limited to things among which the agent can choose. An "overall evaluation" of the set of alternatives implies a ranking of the elements of the set with respect to everything that matters to the agent: desirability, social norms, moral principles, habits—everything relevant to evaluation. Preferences thus imply all-things-considered rankings.

If an agent P has a preference ranking among the actions A, A', A", etc. among which P chooses, then orthodox variants of rational choice theory say that P chooses a feasible action A only if P does not prefer any feasible alternative to A. Unless P is indifferent between A and some alternative A', P chooses a feasible action A if and only if P prefers A to every feasible alternative. To say that P's choices thus track P's preferences says nothing about what it is that P prefers. Though this view of the relationship between choice and preference is consistent with a view of people as hedonists or egoists, it is also consistent with altruism and moral commitment. Given a specification of P's preferences *among the feasible alternative objects of choice themselves*, rational choice theory is trivial. It says only that P will choose among the top-ranked alternatives.

Notice that once preferences among the immediate objects of choice are given, belief drops out of the story. It plays no role in the relationship between action and preferences among the immediate objects of choice. Furthermore, when one can coherently describe a possibility of P choosing between two alternatives, P's all-things-considered ranking of those two alternatives will coincide with how P would choose. People also have preferences among alternatives among which they could not possibly choose. For example, Darcy prefers that Elizabeth accept his proposal of marriage, though whether she accepts or not is her choice and could not possibly be

something that Darcy chooses. Only when the objects of preference are limited to the immediate objects of choice will preference coincide with what economists call "revealed preference."

2. EXPECTED UTILITY THEORY

Though many rational choice theorists reject expected utility theory, expected utility theory is one influential version of rational choice theory, and it is the only version that I shall discuss here. Expected utility theory imposes formal constraints on preferences. Some of these, such as transitivity and independence, are arguably principles of rationality. There seems something irrational about a person's set of preferences if they are intransitive or if preferences among lotteries that differ only in one prize do not match preferences among the prizes. Other axioms, such as completeness, continuity, and reduction of compound lotteries, might be justified as reasonable approximations that permit theorists to separate different issues, but I am not going to argue here for or against these axioms. Cardinal representation theorems show that when all the axioms are satisfied, an agent's preferences can be represented by an index function that is unique up to a positive affine transformation and that has the property that the index attached to a lottery is a sum of the indices attached to the payoffs weighted by their probabilities. These indices are, of course, called "utilities," and misunderstandings stemming from this nomenclature are legion. But with this warning that utility is most definitely *not* itself an object of preference or anything other than an index of the extent to which preferences are satisfied, I shall help myself to the familiar terminology.

Cardinal representation theorems imply that the utility of a lottery for an agent is the mathematical expectation of the utilities of its outcomes: (1) $U(L) = p_1.U(O_1) + p_2.U(O_2) + \ldots + p_n.U(O_n)$. Different treatments offer different interpretations of the probabilities used to calculate the expectation. I shall take them to be the agent's subjective probabilities. If one knows the agent's subjective probabilities and the utilities the agent attaches to the outcomes, then one knows precisely where the lottery stands among the objects of the agent's preferences.

When students first study expected utility, they are inclined to interpret the formula (1) as part of a theory that explains, predicts, or prescribes preferences for lotteries in terms of preferences for outcomes. Knowing U(\$100) and U(\$0), the agent can and ought to calculate that the utility of a gamble of \$100 on a fair coin coming up heads is U(\$100)÷2 + U(\$0)÷2.

Most economists and decision theorists would reject this interpretation of expected utility. It takes expected utility theory to be a theory concerning how agents *form* preferences over lotteries, and it thereby permits a conditional appraisal of whether the agent's preferences among lotteries are rational—that is, an appraisal of the agent's preferences *given* the agent's subjective probabilities and the agent's preferences among the prizes. But decision theorists maintain that expected utilities only *represent* preferences; they do not determine them. Though it is possible to use expected utility theory to guide one's preferences in tricky situations,[1] expected utilities could not be assigned to outcomes in the first place unless agents already had preferences over an infinite set of lotteries.

Rather than regarding expected utility theory as a theory of preference formation, which permits the evaluation of preferences, most decision theorists and economists would maintain that one should regard it merely as *representing* preferences that satisfy its axioms. To the extent that one regards these axioms as requirements of rationality or as reasonable idealizations, ex-

[1] For example, when faced with Allais' problem, in which many people—including even Leonard Savage—are tempted to violate the independence axiom, calculation can save one from making mistakes.

pected utility theory places constraints on sets of preferences. For example, suppose that for some agent, who cares only about money, the utility of a $100 bet on a fair coin landing heads were not the average of U($100) and U($0). In that case, the agent must violate one of the axioms of expected utility theory. If the axioms that are not idealizations are truly conditions on rationality, then either one of the idealizations leads to error, or the agent is irrational. To conform his or her preferences to expected utility theory, the agent must change some preference. But there is no reason for the agent to change the expected utility of the bet rather than to adjust the expected utility of one of the prizes or the subjective probability that the coin lands heads. In other words, there is no theory of the rationality of any specific preference.

The orthodox view is that decision theory in general and economics in particular have nothing to say about where preferences come from or how agents should modify their preferences if they violate the axioms of expected utility. On the contrary, economics focuses on agents whose preferences are complete and already conform to the axioms of expected utility theory. Questions about how agents came to have those preferences and about what therapy should be applied to agents whose preferences are defective are for psychologists or sociologists, not for economists and decision theorists. For example, in the case of standard consumer choice theory, economists suppose that consumers have a complete preference ordering over the commodity space. In deciding how to spend their incomes, consumers calculate how much different bundles of commodities cost so as to identify the set of affordable commodity bundles that best satisfy their preferences.

3. CONSEQUENTIALISM AND PREFERENCE FORMATION

This textbook illustration shows that the orthodox view that preferences are simply givens cannot be the whole story. It misses a complication that arises even in the relatively unproblematic case of consumer choice theory. Although economists take preferences *over the space of commodities* as given, they do not take as given preferences over alternative *choices* or *actions* (commodity bundles purchased). On the contrary, the whole point of consumer choice theory is to show how consumer's preferences among alternative purchases *depend on* incomes, prices, and preferences among commodities. If preferences over purchases were already given, there would be little to explain or to predict. All that would be left for economists to say is that consumers purchase whatever they prefer to purchase.

The explanatory and predictive strategy employed by consumer choice theory is an instance of a more general view, which I shall call "consequentialism." By "consequentialism," I do not mean the ethical view that actions and policies should be evaluated in terms of the goodness of their consequences. I mean instead that an agent's preferences among actions are determined by the agent's subjective probabilities and the agent's evaluations of the results of the actions. The asymmetrical treatment in consumer choice theory of preferences among commodities versus preferences among purchases is an instance of consequentialism. Rational choice theory is not necessarily consequentialist, though its applications often are.

This sense of consequentialism is closely related to Peter Hammond's notion (1988a, 1988b), but it is not the same. My context is broader than the decision trees he discusses, and, more importantly, unlike Hammond, I am taking consequentialism to impose a structure on predictive and explanatory theories of choice: preferences among the objects of choice themselves (like choices) depend on an evaluation of the expected outcomes of choices. This dependence is both causal, as is appropriate in an explanatory theory, and rational, as is appropriate

in a prescriptive theory. Although Hammond may have some explanatory and predictive interests,[2] he is mainly concerned with consequentialism as a rationality condition. Consequentialism can serve as a rationality condition even if it is useless as an account of a structure of explanatory and predictive theorizing.

Although decision theorists need not be consequentialists, they had better have some stories to tell about what preferences for actions depend on. If all that consumer choice theorists could say about why an agent purchased one thing rather than another was that the consumer preferred to make that purchase, rather than relating the action to prices, income, and the consumer's preferences over the commodity space, there would be no Nobel prize in economics. If all that game theorists could say about why individuals play one strategy rather than another is that they prefer that strategy, game-theory texts could be very short indeed.

To avoid misunderstanding, let me emphasize that although the standard theory of consumer choice includes a consequentialist theory of preference formation, it does not include any account of the determinants of preferences *among commodities*.[3] Consequentialist views take preferences over the consequences of alternative actions as given. To preserve the division of labor, whereby questions about how people's tastes are formed and changed are kept out of economics proper, consequentialist economists can insist on a strict asymmetry between preferences among actions (purchases) and preferences among commodities. The latter are givens. Together with prices and incomes, they determine preferences among purchases. There is no reverse dependence of preferences among commodities on preferences among purchases and no theory of the determinants of preferences among commodities. As I have argued elsewhere, this asymmetry in fact breaks down—most clearly in game theory—but the point is not germane to the issues with which this paper is concerned.

A consequentialist view of consumer choice is sensible, and it permits economists to regard the theory of preference formation implicit in consumer choice theory as part of a theory of choice rather than a theory of preference. Preferences among purchases are at most causal intermediaries between consumption choices and the real determinants of those choices, which are incomes, prices, and preferences among commodities. Those who are attracted to revealed preference theory might maintain that preferences among purchases are not distinct from consumption choices and are thus not even intermediaries. So economists can concede that there is a fragment of a theory of preference formation or rational desire embedded within orthodox economics without having to make any particular fuss about it.

Yet the theoretical point remains: If economists want to say more about choice among some set of alternatives than that people choose what they prefer, they need to say something substantial about what influences preferences over the alternatives among which people choose. Furthermore, insofar as they are committed to consequentialism, which relies on an asymmetry between preferences among consequences, which are given, and preferences among actions, which are to be explained by preferences among consequences, economists concede something to the naive student who sees utility theory as accounting for some preferences in terms of others.

---

[2] "The norm . . . is *consequentialist* if it is defined at all decision nodes . . . and specifies consequentially equivalent behaviour in any pair of consequentially equivalent decision trees. Thus does behaviour become explicable merely by its consequences" (1988a, 508).

[3] In this regard the standard theory differs from Kelvin Lancaster's "characteristics" approach (1966) and Gary Becker's "household production function" (1973), both of which derive preferences over commodity bundles from preferences over aspects or consequences of commodities.

4. BELIEF, EVALUATION, AND PREFERENCE

Preferences depend on *beliefs* as well as on raw motivational inclinations. Unlike primitive urges, people's preferences depend on their beliefs concerning the character and consequences of the objects of their preferences. For example, my preference for drinking a glass of clear liquid in front of me rather than pouring it into my car's gas tank depends on whether I believe it is water or gasoline. This means that preferences can be regarded both as the *result* of deliberation and as an *input* into deliberation and that belief enters into choice both as a cause of preference and as a cause of action. Beliefs and preferences can be inputs that influence choices, as they are when my desire for water and my belief that the liquid in front of me is water lead me to drink. Beliefs and preferences can also influence other preferences, as, for example, they do when my preferences among flavors and textures and my aversion to early death coupled with my beliefs about the consequences of consuming different foods determine my preferences between sausage and salad. When the objects of preferences are the alternative choices—that is, *actions*—themselves, then belief plays its full role in the deliberations that result in the preference ranking of actions.

Although economists might reasonably suppose that shoppers cruising the aisles of a grocery store, consulting their shopping lists from time to time as they fill their carts, have already settled their preferences among the various commodities that line the shelves, in other contexts, it can be very difficult for agents to decide what they prefer. For example, in the literature on health-state valuation, economists attempt to assign utilities to health states on the basis of the preferences of respondents. For example, people might be asked to consider a hypothetical choice between 10 remaining years of life in a state with debilitating angina or having a surgery that will restore them to full health for the same ten years with probability p and kill them immediately with probability 1 – p. If death is assigned a utility of zero and full health the utility of one, one can take as a quantitative measure of the respondent's preference for serious angina the value of p for which the respondent is indifferent between the two alternatives. This is a variant of an elicitation technique known as a standard gamble.

As I have argued elsewhere, this way of valuing health states supposes that individual respondents have already (somehow or other) figured out how to value health states. Their values then influence their preferences.[4] Forming such preferences is a cognitively demanding task. As David Feeny, one of the leaders in this field, puts it, "The process of choice in the lottery helps the subject to come to a judgement about just how good or bad the health state being evaluated is" (2002, 517). In the context of a discussion of whose preferences should be consulted, Gold et al. write, " Moreover, those experiencing an acute condition may not be best able *to make well-considered judgments of how severe the state actually would be* in the long run" (1996, 100 [italics added]). Just how bad is it to be unable to run around or to experience pain and shortness of breath when climbing stairs? Preferences among health states do not just "happen" to someone, in the way that preferences among colors or flavors might. People need to figure out what they prefer—to judge how bad or how severe health states are. Just as one can make such judgments rationally or irrationally, so one can form preferences rationally or irrationally.

[4] Other things besides judgments concerning the severity or badness of health states influence preferences. For example, a qualitative study of hypothetical choices like the one described above found that some people said they would choose not to have the surgery, even though they believed that their health prospects would be better, because their responsibilities to helpless dependents ruled out taking such risks (Baker and Robinson 2004).

5. REASON AS THE SLAVE OF THE PASSIONS

Why then do rational choice theorists typically take desires to be "the unmoved movers" for which (in contrast to belief) no theory of rational formation and modification need be given? Why does Hume claim that "reason is, and ought only to be the slave of the passions?" In insisting that reason serves the passions, Hume might only be denying that reason all by itself motivates action or determines desire. Though this view is controversial, I shall not discuss it, because one can grant it and still maintain that there are reasons for and against desires. Just as the empiricist claim that perceptual beliefs are determined in part by sense experience does not preclude an account of the rationality of perceptual beliefs, so the claim that reason does not by itself determine desire does not rule out a theory of the rationality of desire.

Hume apparently makes much stronger claims. He writes,

> Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them. As this opinion may appear somewhat extraordinary, it may not be improper to confirm it by some other considerations.
>
> A passion is an original existence, or, if you will, modification of existence, and contains not any representative quality, which renders it a copy of any other existence or modification. When I am angry, I am actually possest with the passion, and in that emotion have no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high. 'Tis impossible, therefore, that this passion can be opposed by, or be contradictory to truth and reason; since this contradiction consists in the disagreement of ideas, consider'd as copies, with those objects, which they represent.

Hume denies that passions can conflict with reason or truth on the grounds that they are not representations of anything. A rage or a thirst or a yearning may be directed toward an object, but it is just a feeling, not a *representation* of its object as one way or another and thus not correct or incorrect. As Hume puts it in "Of the Standard of Taste," "[A] thousand different sentiments, excited by the same object, are all right; because no sentiment represents what is really in the object. It only marks a certain conformity or relation between the object and the organs or faculties of the mind" (1963, 234). Othello's rage at Desdemona's adultery is thus not in Hume's view only a passion or sentiment. It is a passion accompanied by a judgment—in this case a tragically false judgment.

Hume then goes on:

> What may at first occur on this head, is, that as nothing can be contrary to truth or reason, except what has a reference to it, and as the judgments of our understanding only have this reference, it must follow, that passions can be contrary to reason only so far as they are accompany'd with some judgment or opinion. According to this principle, which is so obvious and natural, `tis only in two senses, that any affection can be call'd unreasonable. First, When a passion, such as hope or fear, grief or joy, despair or security, is founded on the supposition or the existence of objects, which really do not exist. Secondly, When in exerting any passion in action, we chuse means insufficient for the design'd end, and deceive ourselves in our judgment of causes and effects. Where a passion is neither founded on false suppositions, nor chuses means insufficient for the end, the understanding can neither justify nor condemn it. 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. . . . In short, a passion must be accompany'd with some false judgment in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgment. (Hume [1738] 1896, bk. 2, pt. 3, sec. 3)

Although Hume concludes by maintaining that passions are never "properly speaking" unreasonable, he does permit us to call "accompanied passions" unreasonable when they rely on mistaken beliefs about their objects, including especially mistaken causal beliefs. So he can permit us to say (though perhaps not "properly speaking") that it is unreasonable for Harold to be terrified of a stick, which he mistakenly takes to be a snake and mistakenly believes will bite him, and it is unreasonable for Othello to want to murder Desdemona.

When Hume writes—with a good deal of dramatic flair—"'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger," he makes it sound as if rational appraisal of desires is a category mistake, rather like rationally appraising a waterfall or earthquake. And if desires were objectless passions—mere feelings that sometimes accompany object-directed judgments—he would have a case, at least if "reason" is narrowly construed. But, perhaps perversely, I suggest that we read these passages from Hume's *Treatise* as sketching a theory of how desires and preferences can be irrational. How one feels about the destruction of the whole world depends on what one believes about that destruction. Does it involve mass agony, or painless extinction? Is one talking about the end of humanity, the earth, the solar system, or the whole universe? Will this destruction leave one suffering the agonies of Hell for all eternity? It is likely that someone who prefers the destruction of the whole world to the scratching of his or her finger misunderstands what the consequences of the "destruction of the whole world" would be. Having such a desire would typically be irrational or "unreasonable" and by Hume's own standards.[5]

One way to make this point is to borrow some terminology employed in a different context by Amartya Sen and distinguish between "basic" and "non-basic" preferences (1970, chap. 5). "Basic" preferences are preferences that are independent of non-evaluative beliefs. They would appear to instances of what Hume thinks of as passions that are unaccompanied by any "judgment or opinion." Preferences among actions are, however, typically non-basic.

Although basic preferences cannot be evaluated instrumentally and might be said never to be "against *reason*" or "*incorrect,*" they are not beyond evaluation. As Hume argues in "Of the Standard of Taste," even if preferences were mere feelings, without representative content, people's preferences may reveal defects in their capacities, when they rank those things that are bad or unworthy above those that are good. Though not themselves *irrational*, preferences for rotten meat, disreputable company, cruel pleasures, or vulgar sentimentality are *defective*. Such defective preferences limit how well people can live and reflect badly on them.

Non-basic preferences like basic preferences are subject to appraisal as sound or defective. In addition, non-basic preferences may be rational or irrational. Preferences among health states are obviously very far from basic, but so are preferences among everyday objects of consumption, which are sensitive to beliefs about the properties of those objects. Even if the Merck pharmaceutical company had not pulled its anti-inflammatory drug Vioxx off the market, demand for it would have plummeted when news got around that it caused heart attacks and strokes. Changing one's preference ranking of Vioxx and ibuprofen can be rational or irrational.

---

[5] One might reasonably point out that there is a difference between saying that it is irrational for a particular person in a particular context to have the desire and saying that desire itself is irrational, either because its content is irrational or because, regardless of the context, having such a desire is always irrational. But the same distinctions hold among beliefs. Someone who has overindulged may believe irrationally that there is a pink elephant in the room, even though there is nothing irrational about this proposition itself or about believing this proposition in some contexts. It is always irrational to believe in an inconsistent proposition, just as it is irrational to desire that an inconsistent proposition be true.

The Vioxx example seems to bring us back to the naive view that expected utility theory shows how preferences among uncertain prospects derive from subjective probabilities and preferences among the prospect's prizes. One cannot consistently maintain that utility theory merely *represents* an agent's preferences while also endorsing the consequentialist view that an agent's preferences among actions depend and ought rationally to depend on his or her preferences among their possible outcomes and the subjective probabilities of those outcomes.

So the theory of rational choice (in its formulation as expected utility theory) not only needs an account of rational preference formation—at least with respect to preferences among the objects of choice themselves—but many applications of the theory apparently already contains portions of one such account. Why then are rational choice theorists so ready to deny that preferences are subject to rational assessment?

6. WHY "PROSPECTS" ESCAPE RATIONAL APPRAISAL

I have thus far ignored a feature of standard formal developments of rational choice theory that appears to undermine the case for a theory of the rationality of specific preferences. That feature concerns the specification of the *objects* over which preferences are defined. I wrote above of desires for pizza and preferences for one purchase rather than another or for living with angina compared to undergoing a risky surgery. But these are not, strictly speaking, the sorts of things which decision theorists take preferences to rank. John Harsanyi, for example, takes the alternatives over which people have preferences to be "prospects" (1977). In the case of decision making under conditions of certainty, these are "sure prospects"—the results of the alternative actions, where results include not only the causal consequences of the actions, but the actions themselves and all the other features of the world that are causally independent of the actions though nevertheless relevant to how the agent ranks the prospect. In circumstances of risk and circumstances of uncertainty, the objects of preference are risky or uncertain prospects. Formal presentations of expected utility theory do not take the objects of preference to include actions, at least as ordinarily conceived, though one can roughly capture what is meant in ordinary language by saying that an agent prefers taking Vioxx to ibuprofen by saying that the agent prefers the uncertain prospect with which the agent identifies taking Vioxx to the uncertain prospect with which the agent identifies taking ibuprofen. The uncertain prospect that an agent identifies with taking Vioxx would be the set of possible results the agent envisions weighted by the agent's subjective probabilities. Although an agent may identify an action with a prospect, actions are not individuated in the same way that prospects are. Two agents who both take Vioxx apparently perform the same action, but they do not choose the same prospect unless they have the same expectations concerning the consequences.[6]

In taking the objects of preference to be prospects, one builds beliefs into the individuation of the objects of preference and thereby guarantees that preferences among these objects are independent of non-evaluative beliefs. An agent's non-evaluative beliefs determine which prospect

[6] This denial that actions, as ordinarily understood, are objects of preference applies even to versions of decision theory such as Leonard Savage's, which sound as if they take actions to be the primary objects of preference. Savage calls the primary objects of preference "acts," and he defines them as functions "attaching a consequence to each state of the world" (1972, 14). So Savage would identify an individual P's preference for an action such as taking Vioxx with a function attaching consequences such as joint pains, stomach upsets, activity levels, heart attacks, and so forth to states of the world over which P has a subjective probability distribution. Though "Savage acts" are not the same thing as prospects, these characterizations of the object of preferences are interdefinable, and I shall frame my discussion in terms of Harsanyi's "prospects" rather than Savage's "acts."

the agent takes an action to be, but they are irrelevant to the agent's preferences among prospects. Learning that Vioxx causes heart attacks does not change how agents rank the uncertain prospects that they previously identified with taking Vioxx and with taking ibuprofen. It instead leads agents to identify a different and less attractive prospect with taking Vioxx. Preferences and desires are no longer subject to rational appraisal of the sort Hume suggested in the *Treatise*. When Harold is terrified of being bitten by a stick, which he mistakenly takes to be a snake, his mistake lies in his characterization of the state of the world to which he is reacting, not in his fear.

Taking the objects of preference to be prospects does not rule out the possibility that preferences and desires depend on evaluative beliefs and principles of morality or prudence. Like a preference for Jane Austen over Jacqueline Suzanne, preferences over prospects may be sound or defective. Preferences among prospects that are insensitive to human suffering or unaffected by the justice or injustice that prospects involve may be subject to moral criticism. But building beliefs about consequences into prospects rules out the simple ways discussed above of challenging the rationality or correctness of preferences.

In formal presentations of expected utility theory, apparently unreasonable preferences are simply preferences among the wrong prospects. Desires are the "unmoved movers," and the concerns I have raised about the judgments that go into determining preferences or the need for a theory of preference formation must be captured instead by a theory of the rational formation of beliefs about which prospects to identify with actions. What a person needs is not, as Mill maintained, "a reason for his feelings," but a reason for characterizing the object to which the feelings are non-rationally directed one way rather than another. Preferences are impervious to criticism as irrational on the basis of any mistaken non-evaluative beliefs they may presuppose not because, as Hume suggested, passions have no objects, but because all relevant non-evaluative beliefs about the character and consequences of alternatives are already built into the objects of preference. If the beliefs are false, then the actions should be controlled by preferences among different prospects, not by different preferences over the same prospects.

The orthodox framework has its virtues, and one may be able to live with this implication. It does not deny that beliefs are crucial to preferences; it just relocates their importance. Limiting the objects of preference to prospects is one way to take seriously the concerns about the rationality of preference that I raised above. To be sure, one can no longer put things the way I did. In a framework in which preferences do not depend on (non-evaluative) beliefs, one cannot speak about unreasonable or mistaken preferences, which are based on false beliefs. But in insisting on defining the objects of preference as prospects, one is automatically addressing the underlying concerns about the sensitivity of desire to belief, and the identification of actions with prospects would still be subject to rational appraisal.

7. PREFERENCES AMONG ACTIONS ARE STILL SUBJECT TO RATIONAL APPRAISAL

Yet regimenting decision theory in this way misdescribes our evaluative practice, breaks the connections between evaluative judgments and preferences, and may fail to allow for the possibility of certain kinds of mistakes concerning the objects of preferences. The first problem is that the regimentation misdescribes the objects of ordinary preferences. For example, early in *Pride and Prejudice* when Elizabeth would like to dance and Darcy refuses, we take them (and they take themselves) to be disagreeing initially about the evaluation of two alternatives and to have conflicting preferences. But Darcy identifies dancing with a different prospect than does Eliza-

beth. They may not disagree about the ranking of any relevant prospects, and even when they do, what we see are their different attitudes toward dancing not toward prospects. Whether or not to dance are alternatives for which reasons can be given, unlike prospects, which are insulated against prosaic evaluative considerations. Furthermore, to suppose that people, even such thoughtful ones as Elizabeth and Darcy, are able to identify the alternatives they face with well-defined prospects is a tenuous idealization. Though people clearly think about what their choices involve and what their immediate consequences may be, people do not and cannot look very deeply and do not have well-defined subjective probabilities over states of the world or results. Well-defined prospects are rarely the objects of choice.[7] Restricting preferences to prospects seems to get things backwards. It is doubtful whether people have preferences over such complicated things as prospects—while it is obvious that people have preferences over alternative actions.

Of greater concern than the counter-intuitiveness of the idealizations is the way that this construal of the objects of preference disconnects preferences from evaluative judgments. In making preferences impervious to consequentialist considerations concerning values, one makes them mysterious and inevitably winds up understating the role of judgment in the formation of preference. By defining the objects of preference so that preferences are made independent of non-evaluative beliefs—that is, by so defining the objects of preferences that all preferences are basic—one cuts preference off from one important kind of criticism and evaluation. For example, people's judgments about the severity of health states ought to influences their preferences among health states.

People spend much of their lives assessing things, both trivial things like jokes and smells, and serious things like careers and political platforms. People debate and make decisions about the values of things. Some of the considerations influencing such assessments are more like what Hume discusses when he writes about the soundness or defectiveness of taste, but many of these considerations are instrumental or consequential. Sometimes acquiring a trinket may be of value to an agent simply because the agent wants it. More often an agent wants something because he or she judges that it is of value or because getting it or seeking it will be of value. Preferences typically either derive from or reflect value judgments.

Any plausible account of the objects of preferences must permit this dependence of preference on value judgment, and accounts such as Harsanyi's just barely respect this constraint. An account such as Harsanyi's permits preferences among prospects to depend upon agents' evaluations of their features. The only caveat—and it is not a small one—is that changes in non-evaluative beliefs about prospects count as changes in prospects and thus do not affect evaluations of any given prospect.

If one takes the objects of preference to be prospects, then the reasons people give in defense of their preferences have to be sorted into either reasons for taking the objects of preference to be one prospect rather than another or purely evaluative or conative factors governing reactions to features of prospects. Though some Humeans may see this sorting as enforcing clarity and order in our evaluative considerations, this sorting seems to me to make a hash of our evaluative practices. Defining the objects of preference cannot make questions concerning the rationality of preferences disappear. Unless the rephrasing required by the way the objects of

[7] As Richard Jeffrey points out, in Savage's framework "to know what act you are performing you must know exactly how it would turn out in each possible state of nature" (1983, 22).

preference are individuated is simply mistaken, it must provide a way to pose questions about the justification of preferences.

To do this, rational choice theory must allow for the possibility that an agent may be mistaken about the alternatives among which he or she is choosing. The easiest way to do this is to permit actions as well as prospects to be among the objects of preference. For example, consider the following description that Austen offers of Elizabeth's state of mind after she has refused Darcy's proposal and has then read his letter recounting his relations with Wickham,

> Of neither Darcy nor Wickham could she think, without feeling that she had been blind, partial, prejudiced, absurd. . . . After wandering along the lane for two hours, giving way to every variety of thought; re-considering events, determining probabilities, and reconciling herself, as well as she could, to a change so sudden and so important, fatigue, and a recollection of her long absence made her at length return home.

What is at issue here are Elizabeth's appraisals of Wickham and especially Darcy and of her own actions toward Darcy, not the appraisal of prospects. Though Elizabeth's tastes are no doubt evolving throughout the novel, the change in her attitudes that is "so sudden and important" follows from a change in belief rather from a change in basic preferences. Her task is a cognitive one: "reconsidering events" and "determining probabilities."

Preferences among actions, unlike preferences among prospects, depend on fallible beliefs about what prospect an action is. Though Elizabeth has not yet changed her mind about whether she would like to marry Darcy, her attitude toward that action, like almost all preferences, depends on fallible beliefs about the nature and consequences of actions.

8. CONCLUSIONS

Preferences among actions are subject to rational criticism if the identification of actions with prospects is subject to rational criticism. In this way formal versions of rational choice theory can make room for a theory of rational desire as well as a theory of good taste and unite the rational considerations that guide our judgments of value to the account of preference and rational choice. A

REFERENCES

Baker, Rachel, and Angela Robinson. 2004. Responses to standard gambles: Are preferences "well constructed"? *Health Economics* 13:37–48.

Becker, Gary. 1973. On the new theory of consumer behavior. *Swedish Journal of Economics* 75:378–95.

Davidson, Donald. 1963. Actions, reasons, and causes. *Journal of Philosophy* 60:685–700.

Dewey, John. 1929. *The quest for certainty*. Repr., New York: Capricorn Books, 1960.

Elster, Jon. 1983. *Sour grapes: Studies in the subversion of rationality*. Cambridge: Cambridge University Press.

Feeny, David. 2002. The utility approach to assessing population health. In *Summary measures of population health: Concepts, ethics, measurement and applications*, ed. Christopher Murray, Joshua Salomon, Colin Mathers, and Alan Lopez, 515–28. Geneva: World Health Organization.

Gold, Marthe R., Donald L. Patrick, George W. Torrance, Dennis G. Fryback, David C. Hadorn, Mark S. Kamlet, Norman Daniels, and Milton C. Weinstein. 1996. Identifying and valuing outcomes. In *Cost-effectiveness in health and medicine*, ed. Marthe R. Gold, Joanna E. Siegel, Louise B. Russell, and Milton C. Weinstein, 82–134. New York: Oxford University Press.

Hammond, Peter. 1988a. Consequentialism and the independence axiom. In *Risk, decision, and rationality*, ed. B. Munier, 503–16. Dordrecht: Reidel.

———. 1988b. Consequentialist foundations for expected utility. *Theory and Decision* 25:25–78.

Harsanyi, John. 1977. *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.

Hume, David. [1738] 1896. *A treatise of human nature.* Ed. L. A. Selby-Bigge. Oxford: Oxford University Press.

———. 1963. Of the standard of taste. In *Essays: Moral, political, and literary*, 231–55. Oxford: Oxford University Press.

Jeffrey, Richard. 1983. *The logic of decision*. Chicago: University of Chicago Press.

Lancaster, Kelvin. 1966. A new approach to consumer theory. *Journal of Political Economy* 74:132–57.

Mill, John Stuart. 1884. *An examination of Sir William Hamilton's Philosophy*. Vol. 2. New York: Henry Holt.

Savage, Leonard. 1972. *The foundations of statistics*. 2nd ed. New York: Dover.

Sen, Amartya. 1970. *Collective choice and social welfare*. San Francisco: Holden Day.