RATIONAL IRRATIONALITY:
A FRAMEWORK FOR THE NEOCLASSICAL-BEHAVIORAL DEBATE

Bryan Caplan
George Mason University

INTRODUCTION

Economists have characterized beliefs as "rational," if agents satisfy Bayesian probability axioms; or, more strongly, if they also satisfy the rational expectations assumption [Sheshin, 1990; Wittman, 1990]. A diverse body of experimental evidence shows that individuals' beliefs deviate from these standards of rationality [Kahneman, Slovic, and Tversky, 1982; Camerer, 1995; Rabin, 1998]. Critics of these findings argue that the anomalies are suspect because financial incentives were absent or inadequate; people would be more rational—perhaps fully rational—if the monetary rewards were large enough [Harrision, 1992; 1990; 1989; Wittman, 1990; Friedman, 1998; Smith and Walker, 1993]. Defenders of the behavioral perspective reply that anomalies are generally robust to this criticism.

The purpose of the current article is not to resolve this controversy, but to provide a shared framework for debate. I present a model of "rational irrationality" and show that the main positions in the neoclassical-behavioral debate about beliefs are special cases of it. In this model, irrationality is a good like any other, and agents optimize by trading wealth for irrationality. Unless otherwise stated, "irrationality" is interpreted as "deviations from rational expectations," of which deviation from Bayesian axioms is a subset. The central assumption of the model of rational irrationality is that agents perceive the price of irrationality without bias. On some level, they have rational expectations about the consequences of irrationality, even though they typically hold a positive quantity of irrationality in their consumption bundle.

The upshot is that it is not necessary to see the neoclassical and behavioral approaches as two irreconcilable paradigms. The neoclassical-behavioral dispute over beliefs can instead be seen as a disagreement within "normal science" about parameter values; even if researchers cannot agree about their conclusions, at least they are asking the same questions. At the same time, the rational irrationality model does not tautologically define genuine irrationality away; a key falsifiable implication of the model is that compensated demand for irrationality must be decreasing in price. Experimental findings that irrationality increases as monetary incentives rise thus differ in kind from other anomalies and merit special attention [Hogarth and Reder, 1987].
The next section provides an elementary taxonomy of rationality and irrationality to clarify the scope of the current paper’s investigations. Section three presents the simple model of rational irrationality. Section four analyzes four special kinds of wealth/irrationality indifference curves—classical, near-classical, near-behavioral, and behavioral—and their implications for the classical-behavioral debate. The fifth section shows that several forms of wealth-enhancing irrationality can also be understood within the rational irrationality framework. Section six concludes the paper.

A TAXONOMY OF RATIONALITY AND IRRATIONALITY

"Rationality," in economic parlance, is equivocal in at least two ways. Not only has it been ascribed to both preferences and beliefs, but in each domain there is a spectrum of rationality standards, from least to most demanding. In the interests of clarity, this section provides an elementary taxonomy of rationality. By no means intended to be exhaustive, its function is only to delimit the scope of the rational irrationality model’s application.

At the outset, then, it is necessary to distinguish preferences from beliefs [Montgomery, 1996; Aumann, 1987]. An agent’s preferences indicate how he would behave (i.e., what he would choose to do) in all conceivable situations. An agent’s beliefs indicate what probability he would assign to any conceivable situation actually being the case. Figure 1 illustrates this contrast by showing preferences and beliefs as two disjoint sets.

Within the set of preferences, one can then draw the sub-set of “well-ordered and stable” preferences. This is typically the weakest hurdle for preferences to qualify as “rational” [Becker, 1962]. Adding on more restrictive conditions—most commonly, the expected-utility axioms of choice under uncertainty—further shrinks the set of rational preferences [Camerer, 1995]. Of course, both more and less demanding rationality criteria can and have been proposed [Harsanyi and Camerer, 1994].

Next consider the set of beliefs, where quite different senses of “rationality” apply. Typically, the least restrictive definition requires only that beliefs satisfy the familiar Bayesian probability axioms. These set no limits on agents’ prior probabilities; in principle, an agent who is rational in this weak sense could be greedily and systematically mistaken about what the world is actually like. Imposing the more restrictive rational expectations assumption rules out such cases, requiring that agents not only satisfy the Bayesian axioms but also hold unbiased prior probabilities.

The remainder of this paper discusses the rationality of beliefs alone. "Rational" is used interchangeably with "satisfies the rational expectations assumption." The domain of "irrational" beliefs, the shaded region in Figure 1, is the union of (1) non-Bayesian beliefs, and (2) Bayesian beliefs that fail to satisfy the rational expectations assumption. This is not meant to suggest that this is the only form of "irrationality" worth considering—or to ignore the experimental evidence on "irrationality" in other senses of the word. Rather, the current paper focuses on violations from rational expectations because this problem is at once important and tractable. Analyzing behavior without well-ordered preferences is quite difficult; but given well-ordered preferences, analyzing irrational beliefs is—it will be argued—a manageable task.

RATIONAL IRRATIONALITY

Suppose an agent has well-defined preferences over both personal wealth and beliefs; he cares about his wealth, but also has a “bias belief” y that (holding wealth constant) he would most like to believe [Akerlof and Dickens, 1982; Akerlof, 1988; Caplan, 1999a, 1999b]. Wealth should be conceived in broad terms to include not just consumption and portfolio value, but also human capital, health, leisure, and so on. In fact, one could just partition all arguments in the utility function into “beliefs” and “everything else,” and use wealth as a synonym for the latter. These preferences can then be represented with indifference curves in wealth/irrationality space as demonstrated in Figure 2. Wealth is on the x-axis, and the absolute value of the deviation from rational expectations is on the y-axis, an agent with rational expectations con-
suumes zero. The only unusual feature of these indifference curves is that they bend backwards at the bliss belief; an agent is assumed to have some specific belief that he feels attached to, rather than a contrarian desire to be as irrational as possible.

As with utility theory in general, one should not read too much psychological content into this choice model. A hungry person might state that he buys food not because he "wants" it but because he "needs" it. A worker could balk at the suggestion that he "reveals his preference for leisure" when he takes a break after a double shift. An agent who holds systematically biased beliefs that "just seem obvious" to him can be seen in the same light.

In psychological terms, cognitive and motivational biases are different: the motivational depend on the emotions, but the cognitive do not (Nisbett and Ross, 1980). But the rational irrationality model treats them symmetrically for analytical purposes. Still, the introspective experience of the effect of incentives could depend on the type of the bias. For a motivational bias like over-confidence, an agent might describe his response to incentives as: trying to be reasonable, suppressing his emotions, or making an effort to give opposing arguments a fair hearing. For cognitive biases like the availability bias, the same agent might say that incentives make him more likely to doubt his initial intuition, look at aggregate rather than anecdotal evidence, ask for expert advice, or spend extra time researching the question. Modeling cognitive and motivational biases symmetrically does not imply that they are introspectively equivalent.

In most situations of practical interest, systematically biased beliefs have a negative impact on personal wealth. Intuitively, an agent who consistently responds optimally to the way the world is almost certainly fails to respond optimally to the way the world is. As Nisbett and Ross put it, "The costs of willy-nilly distortions in perception are simply too high to make them a curse for the disappointed or threatened perceivers. In general, misperceptions make us less able to remedy the situations that threaten us or give us pain than do accurate perceptions" (1980, 234). Suppose, for example, that a doctor genuinely thinks that 95 percent of the people who test positive for a disease have it, when in fact merely 2 percent do, as in the familiar base rate experiment (Casscells, Schoenberger, and Graboys, 1978). Earnestly acting on this biased perception endangers the doctor's career prospects, risking malpractice suits, dissatisfied customers, and loss of professional reputation. Similarly, an agent who is systematically over-confident, mistaking 80 percent probability for perfect certainty, is likely to make a wide variety of losing bets; a poorly calibrated weather forecaster would be likely to lose his credibility and his audience (Lichtenstein, Fischhoff, and Phillips, 1982). People who overestimate their own abilities, and accordingly bargain for more than they can reasonably expect to get, may reduce their average earnings (Babcock and Loewenstein, 1997).

Agents' wealth/irrationality "budget lines" therefore normally have the familiar negative slope as seen in Figure 3. A wealth/irrationality budget line shows the combinations of wealth and irrationality that are feasible. For the sake of convenience these expected losses are drawn as linear, i.e., proportional to the degree of bias, though this need not be the case. The budget line's intersection with the x-axis shows an individual's wealth assuming he strictly conforms to the rational expectations assumption. The budget line's intersection with the y-axis, in contrast, indicates the level of irrationality necessary to actually drive an actor's wealth down to zero. If wealth is defined broadly enough, the budget line's intersection with the y-axis could be interpreted as the point beyond which more extreme irrationality would be fatal. Rationally irrational agents choose their utility-maximizing combination of wealth and irrationality based on an unbiased judgment about the tradeoffs. The higher the private cost of irrationality becomes, the flatter the budget line gets, and the smaller the optimal purchase of irrationality becomes as Figure 3 shows. The crucial assumption is that agents on some level have rational expectations about the slope of their wealth/irrationality budget line; they perceive the impact of their irrationality on their wealth without bias. This is what differentiates rational irrationality from the competing hypothesis of unqualified irrationality. The point here is not to define unqualified irrationality away, but to lay the groundwork for empirical comparison by spelling out the rational irrationality model's details.

If the private impact of irrationality on wealth were zero, the wealth/irrationality budget line would be vertical, and an optimizing agent would always choose his bliss belief3. This polar case plausibly arises under many circumstances. For example, it is pleasant to believe that your job is socially beneficial, but it unearths how biased beliefs on this point would be costly in terms of material wealth (Klein, 1984). The same applies to many religious and political beliefs: for most people, there are no practical repercussions of doubting the theory of evolution or believing that one's nation is the "best in the world" (Caplan, 1999a).

THE NEOCLASSICAL-BEHAVIORAL CONTINUUM AND ITS PRACTICAL SIGNIFICANCE

"Neoclassical" vs. "Behavioral" Indifference Curves

In standard neoclassical models, agents have no preferences over beliefs. Beliefs are a tool for getting more desirable commodities, not an end in themselves. "Neoclassical" wealth/irrationality preferences can therefore be represented as vertical indifference curves: "neoclassical" agents care only about their wealth, not their opinions.
Within the rational irrationality framework, there is a straightforward way to reformulate the standard neoclassical critique. Suppose that instead of perfectly vertical wealth’s irrationality indifference curves, actors have “near-neoclassical” indifference curves, almost vertical, but curving outwards around $y^* > 0$ as Figure 4 shows. Under most conditions, “near-neoclassical” actors cannot be distinguished from neoclassical actors: Neither is irrational if the price is non-trivial. The latent difference reveals itself only as the price of irrationality ceases zero—as it would in an experiment without monetary incentives. In these choice environments, near-neoclassical actors consume appreciable amounts of irrationality, whereas perfectly rational actors are as rational as ever. If indifference curves were really vertical, then regardless of the absence of incentives, critics of behavioral economics would lack an explanation of systematic experimental choice anomalies. But with “near-neoclassical” preferences, the main critique of behavioral economics is intelligible: people exhibit obvious irrational biases in the absence of material incentives, but at a small positive price this disappears.

The present paper does not try to resolve the behavioral-neoclassical controversy. What the rational irrationality model provides is a common framework for debate. There are not two irreconcilable approaches, but two endpoints on a continuum of possibilities. If the polar neoclassical position is empirically untenable, the polar behavioral position is not the only alternative. Economists with neoclassical priors (i.e., the prior judgment that people’s wealth/irrationality indifference curves are vertical) might instead make the marginal move to the near-neoclassical view as experimental anomalies multiply. Conversely, economists with behavioral priors (i.e., that people’s wealth/irrationality indifference curves are horizontal) need not leap to the neoclassical view if some evidence of price-sensitive irrationality accumulates. They could instead concede only that indifference curves are “near-behavioral”: the consumption of irrationality is price sensitive but positive even at high prices. See Figure 4.

Adopting rational irrationality as a framework can also illuminate Harrison’s [1989] claim that it is more informative to check for deviations from rationality in payoffs rather than beliefs. Harrison notes that in terms of both statistical and economic significance, experimental divergence from full rationality is often large, but due to the flatness of most payoff functions, divergence from income maximization is usually small. How can this argument be related to the neoclassical-behavioral continuum? Suppose that an agent deviates from rationality in neither beliefs nor payoffs. (See Table 1.) This is consistent with both neoclassical and near-neoclassical preferences: the agent might be rational because he is congenitally rational, or because irrationality is costly. Similarly, if an agent deviates from rationality in both beliefs and payoffs, there is evidence for behavioral or near-behavioral preferences: the agent might be irrational because he is congenitally irrational, or because irrationality isn’t costly enough. Observed deviations from rationality in beliefs alone, however, are consistent with everything except for polar neoclassical preferences; only neoclassical preferences are inconsistent with costless irrationality. Finally, note that the fourth box in Table 1 is empty. Deviations from rational beliefs are what drag payoffs below their maximum expected value. It is thus impossible to simultaneously have rational beliefs and diverge from income maximization.
Practical Significance

Even if behavioral anomalies disappear in the face of financial incentives, findings of biased mistakes matter. As Smith and Walker observe, "[t]here are both low-stake and high-stake economic decisions in life, and all are of interest" [1989, 248]. The former can draw near-neoclassical agents to deviate from rational expectations; the latter may provoke near-behavioral agents to set aside their biases [Caplan, 1999a; Camerer, 1987; Frey and Richenberger, 1994]. In addition, sometimes the marginal cost of irrationality is low even though the total stakes involved are high; as Russell and Thaler note, "the more efficient the market, the less discipline the market provides. In a fully arbitrated market, all goods (assets) yield the same characteristics per dollar (returns), thus individuals can choose in any manner without penalty" [1985, 181].

The practical significance of any deviation from neoclassical preferences cannot be confined to lab experiments because this is only one form of low stake decision-making [Kirchgasser and Poenowebus, 1993]. Consider some others: the private cost of systematic misestimates of inflation is small over some ranges [Akerlof and Yellen, 1985]. Contingent valuation surveys overstate the willingness to pay for environmental amenities [Harrison and Krstrom, 1990]. More generally, surveys about issues with expressive value merit suspicion due to the negligible private cost of biased responses [Brennan and Lomasky, 1993; Boulier and Goldfarb, 1998]. Above all else, if rational expectations fail in experiments without adequate incentives, one would expect irrationality to play an important role in democratic elections due to the trivial private consequences of a vote [Caplan, 1999b; Akerlof, 1989].

Wittman specifically tries to contrast hypothetical surveys and elections: "When errors involve little cost (for example, answering a survey question incorrectly in a cognitive-psychology experiment), then little cognition will be employed. When errors involve great cost (for example, purchasing unreliable equipment for the military), then greater cognition will be involved..." [1995, 58]. Considering the enormous unlikeliness that a vote will change an electoral outcome, it is rather the similarity of surveys and elections that stands out. Wittman’s diasonogy confuses private and social costs: in an election, errors may have large social costs, even though they rarely

<table>
<thead>
<tr>
<th>Beliefs</th>
<th>Payoff</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>near-behavioral, behavioral</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>near-neoclassical, near-behavioral, behavioral</td>
<td>nonclassical, near-neoclassical</td>
<td></td>
</tr>
</tbody>
</table>

Will in laboratory surveys. But the private incentive structure of surveys and elections is identical: in both cases, the private cost of errors is effectively zero [Brennan and Lomasky, 1993, 38-41].

Similarly, even if behavioral anomalies are only slightly sensitive to material incentives, it makes a practical difference. Near-behavioral agents will make a wide range of systematic mistakes, but if they encounter the same situation repeatedly the price of irrationality may at last induce them to correct their errors. There are also self-selection effects to consider: if some people are closer to the behavioral pole than others, then people unwilling to abandon their biases even at high prices may switch to activities where their bias makes less difference. This makes the long-run supply of rationality more elastic than evidence from experiments with randomly selected subjects might lead one to expect [Camerer, 1987].

Empirical Implementation

How exactly would empirical workers use this analysis? Two approaches suggest themselves. The first is to test whether (and when) rational irrationality can be rejected in favor of a model of unqualified irrationality. The second is to empirically calibrate the elasticity of irrationality with respect to incentives under various conditions.

Can the rational irrationality model be rejected? The rational irrationality model is consistent with a much broader range of observations than standard neoclassical models, but it cannot be reconciled with a special class of anomalies (income effects aside). Rational irrationality leaves open the possibility that biases fail to decrease as incentives for rationality intensify. But it rules out the possibility that stronger incentives for rationality actually make subjects less rational. This is noteworthy because several extant sources report findings along these lines [Einhorn and Hogarth, 1987; Camerer, 1995]. Einhorn and Hogarth rationalize these results as follows: "Performance, however, depends on both cognition and motivation. Thus, if incentive size can be thought of as analogous to the speed with which one travels in a given direction, cognition determines the direction. Therefore, if incentives are high but cognition fixed, one gets to the wrong place faster." [1987, 43] More straightforwardly, one could say that individuals simply perceive the wrong wealth/irrationality budget constraint.13

What the rational irrationality framework highlights is that such findings are not one more anomaly among many, but a challenge to rationality on the deepest level. As such, they deserve more scrutiny: Harrison [1992, 1995] finds that the marginal incentives in many well-known experiments have been a few pennies or less, and Thaler acknowledges that these findings have not been replicated "at very large stakes" [1987, 96].

One recent and noteworthy study that sheds light on this question is Daniel Friedman’s [1998] study of Monty Hall’s 3-doors paradox. Friedman begins by experimentally demonstrating the paradox’s strength: subjects switch about 30 percent of the time, even though they would switch 100 percent given unbiased probability esti-
mates. He then estimates linear probability models with "switch" (1 if the subject switched, 0 otherwise) as the dependent variable, and tests the sensitivity of the basic results to modifications of the experiment. One variant is to move from regular to "intense" incentives, which increases the ex ante marginal cost of biased estimation five-fold—from $1.10 to $1.50/turn. Friedman uses two different variables to capture the impact: Intense, a dummy variable equal to 1 if a subject had "intense" incentives, and Switchbonus, a continuous variable equal to (cumulative earnings for switching − cumulative earnings for not switching).^6

Friedman reports that in a multiple regression, the coefficient on Intense is actually negative. Its discrete effect is to reduce the "switch" percentage by 9 percentage points, a decline significant at the 5 percent level. [Friedman, 1998, p.540] However, the coefficient on the continuous measure Switchbonus is positive and significant. In combination, the coefficients on Intense and Switchbonus indicate that at least after five turns, intense incentives yield better performance than regular incentives. As Friedman explains, "Since Switchbonus already captures the most relevant positive aspect of intense incentives, the significantly positive Intense coefficient should be thought of as measuring a residual impact" (ibid., 1994).

Friedman's findings are not anomalous, even though he presents strong evidence that considerable irrationality exists. The rational irrationality model moves the analytical spotlight from the "anomaly" in the usual sense of the term (violation of full rationality) to the anomalous response of rationality to stronger material incentives. Now suppose instead that Friedman had found that the coefficients on both Intense and Switchbonus were negative. This would definitely count as empirical evidence against the unlimited applicability of the rational irrationality model, since it would imply that consumption of irrationality rises when the price of irrationality increases.

One potentially important caveat for empirical researchers to take into account is the short-run versus the long-run elasticity of rationality. Since Friedman's discrete and continuous measures of incentives have different signs, strictly speaking, his results predict that at first subjects become less rational when the stakes increase. The positive continuous effect of incentives overtakes the negative discrete effect by turn six. In effect, then, the short-run response of rationality to incentives in the 5-doors game may be negative even though the long-run response is positive.

Divergence between the short- and long-run elasticity of rationality would presumably be far more pronounced if subjects had additional time during which to access a library or the internet. For example, in an alternative experimental design, subjects might play for an hour, receive an hour break of "free time," then return to play for one additional hour. Would higher stakes still have a temporarily perverse effect? Or as incentives rose would players be increasingly inclined to skip lunch and search out a correct explanation [Frey and Eichenberger, 1994]?^7

Calibrating the elasticity of irrationality. A second way for empirical researchers to use the rational irrationality framework is to calibrate the elasticity of irrationality with respect to incentives. Assuming that irrationality declines as the price of irrationality rises, it is still vital to know "how much, under what conditions?" Re-turning to Friedman, one can roughly calculate the elasticity of irrationality in different conditions.

Extrapolating from the probit version of his results, Friedman infers that with intense incentives, subjects would be rational 90 percent of the time after 57 turns. Following the same procedure [Friedman, 1986, p.540] I calculate that with standard incentives, the corresponding rate after 57 turns would be only 35 percent. With non-intense incentives, subjects would reach the same level of rationality only after an estimated 257 turns. In other words, after 57 turns of play, increasing incentives by a factor of 5 increases the percentage of rational responses by a factor of 3.6, and reduces the number of turns required for 90 percent rationality by a factor of 4.5. Focusing on the more immediate effect of incentives, Friedman's results also imply that after 10 turns, the switch rate given intense and non-intense incentives will be 25 and 26 percent respectively. The implied elasticity still has the "right" sign. Yet 10 turns out, magnifying the incentives five times increases the percentage of rational responses by only 14 percent.

Overall, then, Friedman's results suggest that the short-run elasticity of irrationality with respect to incentives is fairly low, but the long-run elasticity is quite high. The reason for discussing the study, though, is not its specific conclusions, but rather to illustrate how the rational irrationality model might redirect empirical research. Within the rational irrationality framework, searching for deviations from full rationality remains useful. In fact, placing irrationality within a consistent theoretical framework would tend to diffuse purely methodological objections against anomalous experimental findings. The rational irrationality framework only demands that evidence of anomalies be accompanied by additional work on their incentive-elasticity under different conditions.

INCENTIVES FOR IRRATIONALITY: POSITIVE ILLUSIONS, MORAL CONSTRAINTS, AND SOCIAL PRESSURE

A simple intuition lies behind the negative slope of the wealth/irrationality budget line: if people respond optimally to the way the world is, not except by pure chance, they will fail to respond optimally to the way the world is. But there are three notable exceptions to this principle. The first case is belief-dependent performance, where an agent's beliefs are actually an argument in his production function; the second, social pressure to deviate from rationality; the third, "self-serving bias" to circumvent fixed moral constraints. In each of these cases, agents actually face an upward-sloping wealth/irrationality budget line shown in Figure 3. As in Figure 3, the budget line is drawn as a straight line for illustrative purposes, but it could easily be non-linear. The steeper the line, the more wealth an agent forgoes if he refuses to delude himself. In effect, agents face a trade-off between a good (wealth), and a bad (irrationality in excess of *y*). To get more of one, they must endure more of the other, much like one might tolerate more pollution in order to produce additional steel.

Note that the positively-sloped budget line reaches a tangency only at some point more irrational than *y*. Even if a person likes being rational (y<0), the positive slope of the budget line provides an incentive for systematic bias in equilibrium. As before,
the responsiveness of beliefs to incentives varies along the neoclassical-behavioral continuum. The closer your indifference curves are to the neoclassical extreme, the more inclined you are to manipulate your views to your material advantage. More neoclassical agents readily "psyche themselves out" to enhance their performance, rationalize to avoid binding moral constraints, and change their opinions as social fashions change. The delusions of those closer to the behavioral pole, in contrast, are relatively stable; their views may be biased, but they are not conveniently interpreting reality to enhance their prospects.

**Belief-Dependent Performance**

Suppose that some of an agent’s beliefs are arguments in that agent's production function. The clearest example is the placebo effect, where individuals' health really improves because they falsely believe they are receiving effective medical treatment. Taylor interestingly notes that "Placebo effects are so powerful that no new treatment or drug can be approved for general use in the practice of medicine unless its effectiveness has been evaluated against that of a placebo" (1989, 113). But there are many non-medical examples too. Psychological studies have also found that frequently individuals with realistic—rather than over-optimistic—probability assessments are more likely to be depressed. With their mood less positive, they are often objectively less successful as a result (Camerer, 1996; Taylor, 1989). Similarly, a student who believes he is likely to do well on a test feels calm and confident, and in consequence probably actually does better.

There is also evidence that the "illusion of control" is on net wealth-enhancing.11 Stress impairs performance, and events believed to be uncontrollable are normally more stressful than equally unpleasant controllable ones. The production function of a person with the "illusion of control" about an unpleasant event is therefore less impaired by stress than the production function of someone immune to the illusion. As Taylor explains:

In terms of the rational irrationality model, agents with any of these forms of belief-dependent performance actually face an upward-sloping budget line—at least over some range as in Figure 5. The flatter the slope of the price line, the greater the impact of irrational beliefs on performance. A person with rational expectations about the probability of success can rationally expect to be less successful and therefore less wealthy than the systematically overconfident. A person with rational expectations about the controllability of stressful events can rationally expect a higher level of stress and a lower level of wealth than a person subject to the illusion of control. As a result, even agents with no intrinsic taste for irrationality (γ = 0) on these topics will decide to be irrational in equilibrium. Perhaps this even explains why overconfidence, the illusion of control, and some other biases, predominate empirically (Lichtenstein, Fischhoff, and Phillips, 1982; Langer, 1982; Taylor, 1989). If over-rating your prospects actually improves your prospects, evolutionary forces tend to weed out tastes for rationality, not irrationality (Cosmides and Tooby, 1996, 84-85; Waldman, 1994).

**Social Pressure**

Belief-dependent performance can matter even for a solitary game against nature, but the scope for wealth-enhancing irrationality expands in multi-player games. The power of positive thinking aside, nature provides no incentive to be irrational, but other humans often do. Suppose people are more likely to hire, befriend, praise, and give political power to those who share their beliefs, and more likely to boycott, shun, denounce, and persecute those who disagree with them (Kuran, 1995; Klein, 1994; Becker, 1974). In this environment, less rationality could easily make you more wealthy, giving rise to another instance of the upward-sloping wealth/irrationality budget line.12 The more intense the social pressure (positive or negative), the flatter the slope of the line (Frey and Eichenberger, 1989).

What is remarkable is that using social pressure, a nucleus of true believers who support some irrational view (that is, with y = 0) can generate positively-sloped budget lines for everyone. Facing social pressure to share in popular illusions, people with no intrinsic inclination towards irrationality (that is, with y = 0) may nevertheless opt to be irrational. A test subject may have no strong feelings about the relative length of two sticks. When surrounded by confederates who all claim that the shorter...
stick is longer, though, test subjects frequently say—and perhaps convince themselves—that they agree with the popular position. On a grander scale, a small clique of committed Communists or believers in the caste system may make dissent so costly that their doctrine wins widespread acceptance [Kuran, 1995]. The "true believers" put pressure on their close associates to share their beliefs, these associates in turn pressure others, and irrationality consequently "trickles down" from activists to the general public.

Moral Constraints and Self-Serving Bias

Rabin [1995] distinguishes between "moral preferences" and "moral constraints." Moral preferences are in the utility function; the more intensely an individual cares about a moral goal, the more he wants at a given price. Moral constraints, on the other hand, shrink one's permissible budget set. People pursue moral goals because they want to, whereas they obey moral constraints because they feel they must. Rabin explains that, "Forgive beliefs, there isn't a big behavioral distinction between these two models of morality," but matters change when people can manipulate their beliefs [1995, 4]. If people have moral preferences, mistaken beliefs misdirect their sincere efforts to do good, making them subjectively worse off. But if they instead face moral constraints, mistaken positive beliefs can expand their budget set and make them subjectively better off.

For example, suppose an agent chooses between two consumptive goods, milk and whiskey, shown in Figure 6. The "moral" budget line shows that combinations the agent can financially afford; the tangency shows the most-preferred bundle, morality aside. If avoiding alcohol were an internalized moral preference, this would be modeled by shifting the agent's indifference curve. The contrary possibility, however, is that an agent avoids alcohol because of a perceived external moral constraint that forbids or discourages its consumption. Unlike the de-internalized moral preference, such a moral constraint would shift the effective budget line in along the x-axis, leaving the agent on a lower indifference curve. Assuming the agent realizes the true alcohol content of whiskey, he has to choose within the "moral" budget set instead of the "amoral" budget set. The knowledge assumption is critical: The more the agent underestimates its alcohol content, the less the moral constraint matters, and the higher his utility.

How can agents make moral constraints less binding? Rabin [1990] focuses on belief manipulation through selective acquisition of information, but biased information processing can in principle be just as effective. In fact, if agents choose their normative as well as positive beliefs, the impact of moral constraints disappears entirely; the "moral budget line" disappears, and they select a point on a higher indifference curve tangent to the "amoral budget line." If beliefs about moral constraints get in the way of self-interest, they could just abandon their beliefs in the relevant binding moral constraints. Suppose, however, that people's normative beliefs are fixed, but positive beliefs are not. Then an agent who maximizes utility subject to personal moral constraints might be better off with unreasonable positive beliefs, or "self-serving bias"[Dahl and Ransom, 1999; Babcock and Loewenstein, 1997].

RATIONAL IRRATIONALITY

FIGURE 6

Moral Constraints

appears to be a binding moral constraint, he rationalizes his way around it, "cheating" on his moral beliefs by manipulating his non-moral beliefs. The more he rationalizes the greater his utility from material goods. For example, the person in Figure 6 might conveniently underestimate the alcohol content of whiskey, moving the "moral budget line" closer to the "amoral budget line."

Note that Figures 5 and 6 illustrate two quite different tradeoffs. Figure 6 shows an ordinary tradeoff between two goods, complicated only by the fact that perceived moral constraints shift the budget line in Figure 5 endogenizes the budget line for Figure 6; it shows that as positive beliefs linked to moral constraints become more biased, perceived moral constraints relax, making a higher level of wealth feasible. The only thing holding back unlimited rationalization is a taste for irrationality. In effect, with choice over beliefs there is an additional margin (wealth/irrationality) to optimize along. The individual with fixed normative beliefs and flexible descriptive beliefs thus faces a positively-sloped wealth/irrationality budget line.

The more onerous the moral constraint is, the flatter is the budget line's slope; the payoff to rationalization increases with the absolute divergence between your morally-unconstrained utility-maximizing action and your morally-constrained utility-maximizing action. The factors Rabin [1990] identifies as mitigating self-serving bias, such as "salience injection," "moral dogmatism," and "moral priming" make equilibrium beliefs more rational by flattening wealth/irrationality indifference curves.

Empirical Implementation

How could empirical researchers use this analysis of wealth-enhancing irrationality? There are two dimensions worth considering. The first is to empirically distinguish between beneficial irrationality and low-cost irrationality. The second is to estimate the incentive-elasticity of wealth-enhancing irrationality. This section discusses the direction empirical work along these lines might take, using overconfidence anoma-
Beneficial irrationality versus low-cost irrationality. Some discussions of wealth-enhancing irrationality (Taylor, 1989) can be unclear whether those with "positive illusions" are (1) better off on balance because they are much happier but only slightly less objectively successful, or (2) better off because they are more objectively successful as well as happier. The rational irrationality model formally distinguishes these two cases: hypothesis (1) is just the standard case shown in Figure 3, while hypothesis (2) is the special case shown in Figure 5. One useful task for empirical researchers would be to determine which of these possibilities is actually the case. For example, does overestimating one's ability to correctly answer general interest questions actually increase the fraction of correct responses by reducing stress?

One way to get at this question, which to my knowledge has not been tried, is to combine standard tests of general knowledge with two sorts of incentives: a reward for a subject's total number of correct responses, and a reward for subjects to accurately guess the number of questions one answered correctly. Designating the number of correct answers as C, the subject's believed number of correct answers C, and rewards/prices p and p, one such compensation formula would be: Payoff = pC - p(C - C). If hypothesis (1) is correct, then raising p should reduce C, but leave C unchanged; if hypothesis (2) is correct, then raising p should reduce both C and C.

Calibrating the elasticity of irrationality. For cases where hypothesis (2) applies, the next task for empirical study is to calibrate the incentive-elasticity of irrationality. The procedure is essentially the same as that in the 3-doors experiment: vary both incentives and other conditions and see how the expected degree of rationality changes. Suppose, for example, that one administers a test of 100 general knowledge questions. One could estimate each subject's "bias belief" for C by setting both p and p equal to 0. The rest of subjects' indifference curves could then be mapped out by varying p and p. For example, one could set (p, p) = (8.30, $0.00) for one group of subjects, and set (p, p) = (8.30, $80) for those remaining. If the average values of V and C for the first group were 55 and 55, and the average values of C and C for the second group were 60 and 65, one could estimate that subjects were willing to increase their bias from 5 percent to 8 percent in exchange for an extra $1.50 worth of income.

The most unusual facet of this problem is that the indifference curves become backward-bending after y. Moreover, while this paper's figures show irrationality as symmetric around y, this is not necessarily the way wealth/irrationality indifference curves actually look. Perhaps beliefs are more malleable (indifference curves are steeper) beyond y than they are between 0 and y, indicating that agents are more stubbornly committed to their wealth-imputing irrationality than they are to their wealth-enhancing irrationality. Thus, calibrating the elasticity of irrationality requires empirical work on both cases.

CONCLUSION

The central message of this paper is that the neoclassical-behavioral debate can be analyzed as an elasticity question. "Pure neoclassical" wealth/irrationality indifference curves are vertical; such agents care only about wealth, not beliefs. "Pure behavioral" indifference curves are horizontal at y; irrationality is constant and unrelated to material incentives. Both endpoints and all of the intermediate cases between them fit coherently into the rational irrationality model. Moreover, this framework is flexible enough to analyze diverse violations of rational expectations. Games of "man versus nature" where more irrationality has a negative impact on private wealth are the most obvious. But it can also be easily applied to games of "man versus himself" (for example belief-dependent performance and fixed moral constraints) and "man versus society" (for example social pressure) where more irrationality makes you materially better off.

While the central purpose of this paper is simply to provide a common framework for debate, it also has two key substantive findings. The first is that experiments where increasing incentives for rationality makes people less rational pose a unique challenge. Such anomalies merit additional attention because they are as incoherent with rational irrationality as they are with rationality. Perhaps the long-run incentive-elasticity of rationality is positive even if the short-run response is negative. Higher stakes in a rigid setting have been found to make biases worse (Thaler, 1987). But what about higher stakes combined with sufficient time to do more research, ask experts, or experiment?

The second conclusion is that even the staunchest critics of the behavioral findings must concede that many people's preferences are not neoclassical, but only near-neoclassical. This seemingly small concession has strong implications. Evolutions claims about the possible scope of irrationality must distinguish between cases where irrationality is privately costly, privately costless, or even privately beneficial; only in the first case are evolutionary arguments for perfectly neoclassical preferences at all compelling. If near-neoclassical preferences are widespread, then behavioral economics provides a distorted picture of market behavior, but ipso facto offers a compelling account of much non-market behavior. Most notably, real-world agents usually lack financial incentives to be rational in politics (Akerlof, 1988; Caplan, 1999a, 1999b). The implication for experimental design: to simulate market conditions, you must use private material rewards, but to simulate political conditions, you must not.1

NOTES

For discussion and useful suggestions I would like to thank Don Boudreaux, Tyler Cowen, Peta Betts, Jim Schneider, Geoffrey Brennan, Bill Dogen, Bill Dobson, Mitch Michels. Ed Loges, J.C. Bradbury, Todd Swanick, David Bernstein, Robin Hanson, Dae Klin, Alex Taha, Niki Tyea, Timur Kuran, Rie Heiner, Roger Congleton, Fab Rejas, David Levy, Kenneth Kolfard, Daniel Freedman, two anonymous referees, seminar participants at George Mason, participants at the Public Choice Outrech seminar and the Public Choice Society meetings, and members of my Armchair Economists' labore. Gisele Silva and Eric Cumpston provided excellent research assistance. The standard disclaimer applies.

1. In contrast, they have characterized preferences as "rational" if they are well-ordered and stable; or, more strongly, if they also satisfy the expected-utility axiom of choice under uncertainty (Scheier, 1092, Camerer, 1995). The current paper focuses solely on the rationality of beliefs. The second section of this paper presents a basic taxonomy of irrationality.
RATIONAL IRRATIONALITY

14. Switchboards should not be interpreted as a simple practice effect, because the set of independent variables also includes a trend variable.

15. On a basic Altavista search of "do doors Premier," the first three hits explain the correct answer. Even in a search of "switch play 3," one of the first ten hits offers a correct explanation.

16. For the part of the experiment with both intense and standard incongruents, subjects played either 22 or 15 turns.

17. The "illusion of control" is defined as an expectancy of personal success probability (appropriately higher than the objective probability would warrant). (Langer 1968, 1968).

18. More generally, it consists in holding change events as controllable.

19. A second possibility is to proceed from the premise that people are more inclined to resist change when they expect that it will be beneficial to them, rather than when it is likely to be harmful to them, even when the two events are identical, as in the case of the "door Premier" examples.

20. And not even in every case, as Akreel and Yellen (1988) show, or as Russell and Thaler’s (1988) findings on consumers’ detergent choices suggest.

21. Rationality in politics is in effect a public good, which, as numerous experiments show, will tend to be under-supplied even if it is not entirely absent (Loury, 1985).

REFERENCES


