

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Advancing Beyond "Advances in Behavioral Economics"

Citation

Fudenberg, Drew. 2006. Advancing beyond "Advances in Behavioral Economics". Journal of Economic Literature 44(3): 694-711.

Published version

https://doi.org/10.1257/jel.44.3.694

Link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:3208222

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

Advancing Beyond "*Advances in Behavioral Economics*[1]"

Drew Fudenberg[2]

First draft: October 12, 2005
This draft: November 1, 2005

"The difference between economics and psychology is that we psychologists never start our talks with assumptions that are wrong."
"That's because you psychologists never make any assumptions at all."
Apocryphal

1. Introduction

In recent years, behavioral economics has changed from niche topic to one that is well represented in all of the major journals. For economists who are wondering what all the fuss is about, *Advances in Behavioral Economics* is an excellent introduction to the field. *Advances* reprints a selection of the "greatest behavioral hits" of the 90's, eighteen papers I would characterize as new research and seven surveys that were previously published in books or in review journals.[3] In addition, the editors contributed a substantial overview chapter, which outlines the contents of the book, argues for the importance of behavioral economics for economics as a whole, and speculates about promising new directions of research. The book's wide coverage and its well-articulated arguments for the field make it a valuable reference and teaching aide, and the included surveys will help newcomers catch up with what is now a very large literature.

The book's scope and arguments also make it a convenient platform for evaluating and critiquing the field as a whole, and that will be the main focus of this essay. I will not try to survey the entire field, or even all of the papers in *Advances;* the first task would require a book and the second at least a monograph. Instead, my comments will emphasize the ways the field could improve, as opposed to its successes so far. But first I should say a few words about the book itself.

Other than the new chapter 1, which was written by the editors, the rest of the book is unaltered reprints of previous papers. Part 2, "Basic Topics," includes surveys, experimental results, and theoretical models; the Table of Contents groups these chapters by topic (e.g. "Fairness and Social Preference" or "Reference Dependence and Loss Aversion".) Part 3, "Applications," consists mostly of analyses of field data, but also includes a survey that focuses on data from experiments. At 740 pages, there is far too much material here for the book to be read straight through, and many readers may prefer to read all of the chapters that deal with a particular topic instead of reading the chapters in order. For example, Chapter 6 ("Time Discounting and Time Preference: A Critical

Review", by Shane Frederick, George Loewenstein, and Ted O'Donoghue) and Chapter 7 "(Doing it Now or Doing it Later", by Ted O'Donoghue and Matthew Rabin) are under the heading "Intertemporal Choice," while Laibson's "Golden Eggs and Hyperbolic Discounting" is in the "Macroeconomics and Savings" section of Part 3; a reader interested in the behavioral economics of temptation and self-control might want to read all three of these chapters before reading about e.g. social preferences. For that reason, it would have been nice to have a page at the start of each topic group in Part 2 that explained the relationship between the Part 2 papers and those in Part 3. It would also been nice to have a description of the relationship between the papers in the same section, as when one paper is a survey that discusses the following one.

More ambitiously, it would have been very helpful for the book to have some new material that commented on the reprinted papers. What new work has either reinforced or questioned the paper's conclusions? Do the authors endorse all of the claims of the included papers with equal confidence, or do some of them seem more convincing than others? Which claims seem plausible but need further investigation? This sort of commentary is unusual in collections of papers, but it is helpful in textbooks and surveys.

2. Progress to Date

One accomplishment of behavioral economics has been simply to draw economists' attention to a range of facts and issues that seem important, such as anchoring and base-rate neglect in probability judgments (Amos Tversky and Daniel Kahneman [1974]) cognitive dissonance (George Akerlof and William Dickens [1982]), and the endowment effect (Jack Knetsch [1989].) A second accomplishment has been to develop formal models that generate and explain these regularities, and can be incorporated into larger models of markets or multi-agent experiments. This second, more theoretical, agenda is the more likely to have a deep and lasting impact on the rest of economics, so improving it is the key to further advances in the field.

Like the editors of *Advances,* I think that theories (both in economics and more generally) should be judged by George Stigler's [1965] three criteria: accuracy of predictions, generality, and tractability. The standard model of individual behavior does very well in terms of generality and tractability, but behavioral economics has helped

highlight some areas where the standard model's predictions are sufficiently wide of the mark that changes are valuable. The challenge for the field is to generate more accurate predictions without sacrificing too much on the other two of Stigler's criteria.

Advances focuses on three of the behavioral models that have been the most successful in this regard, namely models of loss aversion in the spirit of prospect theory (Chapters 4 and 5) , the quasi-hyperbolic model of intertemporal choice (as in David Laibson [1997] and Chapter 15) , and the Enrst Fehr and Klaus Schmidt [1999 and chapter 9] model of social preferences. Researchers continue to look for more general and elegant ways to capture the same sets of facts,[4] [5] [6] and in my opinion it is too early to decide that any of these models should be accepted as canonical. However, even if one is not convinced by the specifics of the various models, they have been used to explain a wide range of observations, which at the least suggests that there are significant empirical regularities in the areas the models address. In addition, all of these models are tractable enough to be embedded in richer and more complex settings, so that behavioral economists can and will adapt the models to explain yet more facts. These models, and the diverse set of experimental and field data in *Advances,* definitely justify Chapter 1's claim that it is "unwise and inefficient to do economics without paying *some* attention to good psychology."

Since the publication of *Advances,* the field has continued to evolve and progress, which further validates that position. The most notable new development is the increased use of data from neural imaging, for example by Daniel McClure et al [2004] and Dominique deQuervain et al [2004. This is intriguing work, but its implications are subtle and perhaps more ambiguous than the research papers acknowledge; I discuss a few of the possible complications involved in Section 4.

Another important development is the increased work on the implications of "psychologically-based" preferences, and non-standard preferences more generally, for market outcomes. *A priori,* one might expect that these preferences will have the clearest impact in monopoly markers, and that market competition would in some cases limit both the impact of "behavioral" agents on prices and the extent to which these agents are exploited by others. These theoretical questions are explored in recent work by e.g. Daniel Benjamin [2005], Stefano DellaVigna and Ulrike Malmendier [2004],

Edward Glaeser [2005], Laibson and Leeat Yariv [2005], Karl Schlag [2005], Jesse Shapiro [2005], and Rani Spiegler [2005].

Finally, I should commend the recent literature that looks for evidence that the various sorts of behavioral preferences that subjects exhibit in laboratory experiments are actually observed in field data. Recent work by Niva Ashraf, Dean Karlan, and Wesley Yin [2005] and Stefano DellaVigna and Ulrike Malmendier [2005] shows that agents are willing to pay a premium to reduce their options in some real-world decision problems. Oriana Bandiera, Ivan Barankay, and Imran Rasul [2005] find evidence for social preferences in the difference between worker effort under relative incentives and under piece rates, and Marianne Bertrand et al [2005] show that the way ads for bank loans are "framed" can have a substantial impact on market demand.

2. Advancing Beyond *Advances*

The work described above is very interesting, and it is clear that much more can be learned by further research along these lines. However, unless the insights and stylized facts obtained so far are related to a small number of models of individual behavior, with some guidelines for when each model should be expected to apply, behavioral economics may remain a distinct field with its own methodology. Chapter 1 articulates a larger ambition: The hope that some sorts of "behavioral" models will lose "their special semantic status" (that is, the adjective "behavioral") and become more widely taught and used. If that happens, the field will have advanced beyond "Behavioral Economics," and the sequel to this volume won't need to be called "More Advances in Behavioral Economics" or "Advances volume 2. From the editors' perspective, the ideal situation might be one where the sequel could be called "Modern Decision Theory" or "Models of Consumer Choice."

To achieve this goal of becoming "normal economics," the field will need to confront several issues; my comments here are intended to help it do so. Because of the context of this essay, my comments are phrased as criticisms of *Advances,* and mostly as criticisms of Chapter 1, but that should not obscure the respect that I have for the editors, or for the field as a whole. Indeed, Chapter 1 acknowledges most of what I see as the key

challenges facing the field. Still, from my perspective the chapter probably understates the difficulties these challenges pose, so one of my goals is to better highlight the potential problems. Also, Chapter 1 doesn't really acknowledge the extent to which the various parts of the book tend to focus on only one or two of these issues and ignore all of the others. This may be partly justified by the "different tools for each situation" approach (see below), but I think that it would be useful for more behavioral economists to think about how the various "behavioral critiques" fit together.

A Set of Assumptions should be evaluated as whole.

As Chapter 1 declares, the standard approach in developing theories in behavioral economics is to "[…] modify one of two assumptions of the standard theory in the direction of greater psychological realism." This approach presumes that since economists found a given set of assumptions useful, the only issue in changing one of them is whether the new assumption on its own seems reasonable given the motivating facts. However, this approach overlooks the fact that the factors that support one modification of the standard model may be correlated with factors that argue for further modifications. Thus, modifying one or two assumptions and leaving the rest unchanged may lead to a logically consistent alternative model, but one whose domain of application is unclear or non-existent. I illustrate this argument later in this section in discussing some applications of equilibrium analysis in behavioral economics, and also to some models of temptation and self-control, but these are not meant to be an exhaustive list. More generally, my point is that behavioral economics could be improved by adding a second step to the standard model-creation process described above: After modifying one or two of the standard assumptions, the modeler should consider whether the other assumptions are likely to be at least approximately correct in the situations the model is intended to describe, or whether the initial modifications suggest that other assumptions should be modified as well.

Choice Overload in Modeling Choice

The fact that behavioral economists do not typically examine the domain where all of their assumptions might simultaneously apply has probably contributed to one

problem with the literature in this field: There are too many behavioral theories, most of which have too few applications. Chapter 1 of *Advances* addresses this concern, and replies roughly that it isn't really a problem, as models correspond to tools, and a bigger toolkit is better. I did not find that response very convincing. It is true that some economic models assume risk-neutrality and others assume risk aversion, some models assume selfish preferences and others add a bequest motive, and so on. However, the current state of behavioral economics offers far too many tools, and too little guidance about when to use each one; without that guidance, a bigger toolkit need not help.[7]

As an example of the problem, consider the question of how to model mistakes in inference. I know that agents do make various types of mistakes in this task, but I don't know when the various mistakes are likely to be either more common or more significant. When I want to incorporate mistakes in inference into a model, should I follow the "confirmatory bias" models of Rabin and Schrag [1999] and Yaariv [2005], and assume that agents miscode evidence that their prior says is unlikely? Or Rabin [2002], and assume that agents update as Bayesians but treat independent draws as draws with replacement? Or Nicolas Barberis et al [1998] and assume that agents mistakenly think they see trends in i.i.d. data? Or some combination of these? Note that the problem here is different than deciding when to include e.g. bequest motives in a model: Bequest motives matter for lifetime savings, and hence influence the shadow value of wealth, but are unlikely to be relevant for many consumption decisions, while all of the inference mistakes described above are potentially relevant in a given inference problem.

What should behavioral theories do?

The proliferation of theories also raises the question of what the theories are trying to do. Since psychology papers tend to be less formal than economists are used to, one useful but minor role of behavioral theories is simply to give a precise statement of the chosen behavioral regularity. A second role is to exhibit a set of assumptions that can generate the specified behavior, be it the endowment effect or the law of small numbers. As an economic theorist, I don't find either of these objectives very satisfying, and certainly neither is grounds for teaching the model in question in first-year microeconomic theory, any more than we teach the theory of various econometrically-

convenient demand systems; these and other worthwhile models that have fairly special domains of applicability are taught in the various field courses. This is not to say that I disagree with the authors' goal of integrating some aspects of behavioral economics into first-year classes, and indeed I already integrate some behavioral facts (as opposed to behavioral theories) to illustrate the limitations of standard assumptions such as transitive preferences, independence of irrelevant alternatives, and common knowledge of rationality. In addition, some behavioral models (such as that of Fehr and Schmidt) fit easily into the standard framework, and can be used as examples or homework problems. However, before behavioral theory can be integrated into mainstream economics, the many assumptions that underlie its various models should eventually be reduced to the implications of a smaller set of more primitive assumptions. [8]

In the long run, we could hope to derive behavioral phenomena such as mental accounting and confirmation bias from such basic properties such as bounded cognition (with a cost in both reaction time and resources for more accurate results) and the modular structure of the brain. We are probably a long way from being able to do this, but in the meantime, behavioral economists and (and economic theorists!) should devote more effort to synthesizing existing models and developing more general ones, and less effort to modeling yet another particular behavioral observation. To that end, the three areas I mentioned above- prospect theory/loss aversion, social preferences, and quasi-hyperbolic preferences- should be held up as examples for emulation. One or two more such "multi-use" theories will be far more valuable for behavioral economics and economics as a whole than any number of specail-purpose models whose main function is to formalize observations from the psychology literature.

Context and Cues

As Chapter 1 says, behavioral economics has emphasized the malleability and context-dependence of preferences and behavior. Reinforcing this point, Chapter 17 on "Money Illusion" (Eldar Shafir, Peter Diamond, and Amos Tversky [1999]) focuses on a empirical examples of a specific sort of framing effect, namely the tendency for decisions to depend on absolute as well as relative price level. Unfortunately, framing and context are very difficult to capture in formal models, and are ignored in most of the

more formal papers in the field and in this book. This is true in particular of the papers on the topics of social preferences, even though these phenomena have been shown to be very susceptible to framing. For example, the weight experimental subjects give to other subject's payoffs can be altered by pre-interaction "speeches," and also by manipulations such as choosing the "dictator" in an ultimatum game by the results of a quiz. Moreover, in some settings, such as wage negotiations, one or both of the parties involved may be both willing and able to manipulate the way the issue is framed.[9] This makes the need to a model of frame-determination all the more important.

Another aspect of the malleability of preferences is the way that people to some extent view money rewards as being "as immediate and tempting" as an immediate utility payoff. Since people cannot literally consume currency, why do they act as if current monetary rewards are tempting? In brief ,it seems that the "impulsive" or short-run self treats money as a cue for an immediate reward even though the only real consequence of earning money is in the future.[10] This subjective equivalence of money and immediate gratification is important for the interpretation of empirical studies that show people exhibit "preference reversal" about the timing of payoffs, as these studies almost always examine monetary payoffs and not consumption choices. [11] It is also important for understanding the fact that human subjects exhibit a paradoxically large amount of risk aversion to small money gambles.[12] However, the tempting nature of money leaves open the questions of exactly which financial rewards we should expect agents to view as tempting, and what other sorts of deferred rewards will be treated in the same way. Is money tempting if it will be received later today? What about money to be received tomorrow afternoon? When are people tempted by other sorts of "vouchers" for delayed consumption?[13] If it is costly or difficult to postpone gratification from today until next week, is it equally costly to postpone it until tonight? How does the answer depend on exactly what sort of gratification is at issue, and on where the decision is made?

I do not know the answers to these questions, but I would like to suggest that the underlying mechanism (which is presumably some form of associative learning) may also underlie some types of framing effects. That is, an agent might be more tempted by one verbal description of a financial reward than another, even though they offer exactly the same sets of consumption possibilities, and the difference in these temptations may be

linked to how the agent has learned to respond to various cues. (Note that people are not born with this sort of response to money; it is acquired as they learn how money is used) In any event, while the study of learning and conditioning may eventually lead to useful models of frames, cues, and "mental accounts" (see Chapter 3), for the time being they are a crucial but unexplained part of many behavioral analyses.

Equilibrium: When and Why

The next set of issues is of importance both for behavioral and "regular" economics. First, when is equilibrium analysis likely to be a good approximation of observed outcomes? Second, when it is not, what sort of models should be used to predict behavior outside of the lab? Game theorists have long understood that equilibrium analysis is unlikely to be a good predictor of the outcome the first time people play an unfamiliar game, and I think it is uncontroversial that some aspects of economic life are best described by non-equilibrium play.[14] However, I argue below that the answer to the first question is less obvious than Chapter 1 suggests, and so far behavioral economists have been reliant on equilibrium analysis for developing models of market outcomes. Of course, equilibrium is also the standard assumption in "non-behavioral" economic applications of game theory, but the extensive discussion of models of initial period, non-equilibrium, play in chapters 12 and 13 of *Advances* (Vincent Crawford [1997] and Colin Camerer [2004] respectively) , and the book's inclusion of this topic as part of "behavioral economics," highlights the need for this material to be synthesized into the rest of the field.

A brief review of the literature will help explain why it is not always obvious when the outcome of a game will approximate an equilibrium. There are extensive theoretical and experimental literatures on "learning in games," based on the idea that equilibrium can arise as the result of a non-equilibrium process of learning, imitation, or adaptation.[15] The former investigates the long-run properties of various learning models, comparing their performance (from the viewpoint of individual agents) and convergence properties (which processes converge to equilibrium in which classes of games), while the latter tries to distinguish between learning models on the basis of experimental data. Most of the formal models of learning in games, and most game-theory experiments, rely

on the idea that agents play a game repeatedly against different opponents, in order to abstract from repeated game effects; this necessarily implies that agents use their experience with past opponents to guide their actions in the current game. It is tempting to conclude from these models and experiments that equilibrium analysis almost never applies in the field, as agents rarely play exactly the same game a great many times. But as argued in Fudenberg and David Kreps [1993] and Kreps [1990], any sort of learning involves extrapolation from past observations to settings that are deemed (implicitly or explicitly) to be similar, so what matters is how often agents have played "similar" games. In addition, in field settings, unlike the lab, there are additional sources of information beyond direct experience: Agents may talk engage in "social learning" by asking the opinions and advice of friends, parents, and neighbors, and in some cases (such as retirement savings) they can also consult books, magazines, and outside experts.[16] The possibility of non-equilibrium social learning does not mean that society effectively pools all information and ends up at an equilibrium, but it does mean that the applicability of equilibrium analysis to most field data is an empirical question that can't be resolved by *a priori* arguments.

Unfortunately, once one leaves the controlled laboratory environment it is not clear how to identify equilibrium vs. non-equilibrium play. If one is certain that payoffs are constant over time, then any movement in play at all shows that agents are not playing a static equilibrium, but this leaves open both the possibility that payoff functions vary and that play corresponds to the equilibrium of some dynamic game. So what is needed is a plausible set of identifying restrictions on the nature of payoffs and strategies, and a model of non-equilibrium play that can be econometrically implemented when the actual payoff functions of the players are unknown to the analyst.[17] Until something like this is done, the implications of chapters 12 and 13 for field data will be difficult to determine, and it will be difficult to incorporate non-equilibrium reasoning with the rest of behavioral economics.

Equilibrium Analysis in Behavioral Economics

Although the status of equilibrium analysis when agents are "rational" is a problem for all of economics, the assumption of equilibrium and the choice of the

appropriate equilibrium concept is even more problematic in some behavioral models. As it relates to equilibrium, the "change one assumption" approach to behavioral economics is to assume agents have a specific form of cognitive imperfection and then adopt a version of Nash equilibrium that is as close as possible in form to the usual one. The problem with this approach is that, as noted above, the usual rationales for Nash equilibrium (at least in laboratory experiments) rely on unbiased learning by the agents. If, as in Rabin and Schrag [1999], agents suffer from confirmation bias in learning about the distribution of chance moves, then it seems likely they would suffer from a similar bias in learning about opponents' play. If they do, then models of non-equilibrium adjustment based on learning will not typically lead to Nash equilibrium.

A similar concern arise in evaluating Eric Eyster and Mathew Rabin's [2005] concept of "cursed equilibrium." Experimental evidence shows in common-value auctions agents overbid and are thus subject to the "winner's curse." Eyster and Rabin argue that this fact, and related errors in other incomplete-information games, can be explained by the concept of ψ-cursed equilibrium. In this equilibrium, players have correct beliefs about the joint distribution of types, and also have correct beliefs about the aggregate distribution of opponents' play, conditional on each of their own types. However, instead of playing the best response to the actual opponents' strategies, each player chooses the action that is the best response to a convex combination of the actual strategies and the aggregate distribution, with weight ψ on the aggregate distribution. In the "fully cursed" case where $\psi = 1$, agents completely ignore the correlation between other players' actions and their types; this corresponds to the outcome of a learning model in which agents observe opponents' actions but neither the opponents' types nor their own payoffs.[18] The case $\psi = 0$ is even easier to explain, as it corresponds to the usual Bayesian-Nash equilibrium. However, it is hard to imagine a reasonable learning process that leads to the intermediate cases. It is true that in some cases intermediate values of ψ fit better than either extreme, but this is not evidence that the model is a good approximation of what is really going on. In addition, the fact that the amount of "cursedness" typically declines as subjects become more experienced suggests that the curse, while real, is not an equilibrium phenomenon.[19]

A related problem with Bayesian equilibrium arises in Roland Benabou and Jean Tirole's [2003] model of a Bayesian equilibrium between various "selves," where each self knows the distribution of the other selves' possible "types", and also their equilibrium strategies. The situation here differs in that each self has correct beliefs, so there is not an inherent conflict with the assumption of equilibrium. Rather, the problem is that for non-equilibrium learning to lead to the Benabou-Tirole equilibrium, the player's type(s) would have to be independently and identically distributed across repetition. If, as seems more reasonable in this setting, the player's "type" (e.g. ability or self-control ability) is fixed once and for all, then learning will not lead to the Bayesian equilibrium that is analyzed by Benabou and Tirole, but to the Nash equilibrium for the game where the payoff function is known.

The O'Donoghue and Rabin $(\beta, \hat{\beta})$ model of time preference [1999, 2001] has complete information, in the sense that each of the "selves" is certain of the payoff functions of the others, but it suffers from a related problem. In this model, at each date *t* the agent values utilities at future dates τ at $\beta\delta^{\tau-t}$ current utils, but forecasts that his future play corresponds to the multiple-selves equilibrium of the standard (or "sophisticated") quasi-hyperbolic model in which the selves playing at each date *t* evaluate future utilities at $\hat{\beta}\delta^{\tau-t}$ utils, and this is common knowledge between the selves. The model is motivated by evidence that people sometimes misperceive their own behavior, but it is not clear how anyone would come to make this particular form of mistaken forecast, and once again a non-equilibrium model might be a better match for the facts the model is trying to explain. Of course, one reason for the use of equilibrium models here is simply the lack of a standard off-the-shelf alternative, and despite their flaws these "faux-equilibrium" models have been useful in showing that these sorts of behavior can be modeled and analyzed, instead of merely noted and then ignored. Still, I think that behavioral economics would be well served by concerted attempts to provide learning-theoretic (or any other) foundations for its equilibrium concepts. At the least, this process might provide a better understanding of when the currently-used concepts apply, but I expect that a serious effort to find foundations will typically end up

suggesting somewhat different, and more accurate, solution concepts than the ones I have criticized above.

Models of Temptation and Self-Control

Equilibrium analysis is not the only area of behavioral economics where the usual change-one-assumption approach overlooks the question of how the entire set of assumptions fits together. As a second example, consider the question of how to model the idea that agents know they have a self-control problem, and so can be willing to pay a premium to reduce their own future choices. The two leading models of this idea are quasi-hyperbolic preferences, as in Laibson [1997 and Chapter 15], and the Gul and Pesendorfer [2001] axioms and corresponding representation. Each of these approaches either implicitly (Laibson) or explicitly (Gul and Pesendorfer) imposes a form of the classical "independence axiom" on choices over menus of actions. However, if agents will face a self-control problem in choosing an item from a menu, it is not obvious that the independence axiom should apply. Moreover, the independence axiom should be expected to generally fail if agents must take some sort of self-control action, just as it does when agents must commit to some of their consumption decisions before knowing the outcome of a wealth lottery.[20] Thus the independence axiom is less compelling here on *a priori* grounds than in the standard model, where it has a normative justification. In addition, Fudenberg and Levine [2005] argue that evidence that self control is a scarce resource (Muraven and Baumeister [2000], Muraven, Tice, and Baumeister [1998]) and is impaired by cognitive load (Shiv and Fedorhikin [1999]. Ward and Mann [2000]) supports models of self-control that are not consistent with the form of the independence axiom that the quasi-hyperbolic and Gul-Pesendorfer frameworks assume.

Once again, the models I am criticizing have been useful and important, which is not by accident- there is little point in explaining how to improve on models that are widely viewed as flawed or uninteresting.

Bounded Rationality

In recent years, behavioral economics has evolved more or less independently of the literature on "bounded rationality."[21] It is hard to give a precise definition of bounded

rationality, or to draw a sharp line between it and behavioral economics, but bounded rationality papers typically suppose that some agents (consumers, firms, or both) use exogenous "rules of thumb," and then derive the consequences. In principle, it would be nicer to derive these exogenous rules from a small set of fairly standard assumptions, and one might hope that behavioral economics could eventually do so. Even when a formal derivation isn't possible, one might feel the conjectured rule is more plausible if it can be shown to be rooted in psychological observations that apply more generally. However, the psychology literature is large, and many of its claims are imprecise, so simply finding a concept or claim in the psychology literature that is consistent with the conjecture may not make the rule any more convincing. This is particularly true when the psychology literature doesn't provide sharp restrictions on just when one should expect to see the behavior in question.

As an example, consider the fact that in some settings some consumers seem to use ignore relevant product characteristics, such as shipping costs (Hossain and John Morgan [2004]) or the cost of ink for printers (Xavier Gabaix and Laibson [2005].) Moreover, firms in some markets act as if they are aware of the consumers' tendency to ignore characteristics, and reinforce it by making the relevant information hard to find (Glenn Ellison and Sarah Fisher Ellison [2005].) We don't expect consumers to neglect all sorts of information, but we also don't seem to have a useful theory of just what will be ignored in which situations, so for the time being it seems better to simply make this neglect an ad-hoc assumption.[22] Similarly, in some situations people appear to use simple heuristics such as "pick a product with probability equal to its market share" (Smallwood and Conlisk [1978]) or "identify the average payoff or utility of a product with its value in a small sample (Ellison and Fudenberg [1995], Spiegler [2004].) One long-term goal for behavioral economics is to incorporate more of these ad-hoc rules into its formal framework. This will require a bit of a methodological shift for the field, as it involves looking for inspiration to the older economics literature in addition to psychology.

4. Interviews and Inner-views

Behavioral economists are beginning to use two sorts of data that lie outside the traditional scope of economics, namely questionnaires about mental states (e.g. "How happy are you right now?" " How happy would you be if x occurred ?") and data such as neural imaging on physiological processes inside the brain. On a classical revealed-preference view, neither sort of data is of interest, but I believe that this data can indeed be of use, provided it is interpreted correctly.

Neuro-economics: Do economists need (data about) brains?.

Neural imaging work so far is suggestive of how certain sorts of decisions are made. Even if one takes the view that the only goal of economics is predicting behavior on the basis of "external" variables, having a model that is in better accord with the underlying structure of the brain can be valuable, as it may lead to more accurate out-of-sample predictions.[23]

However, the interpretation of neural imaging and neurobiological data can be difficult and subtle, especially when one tries to use imaging data to resolve debates, as opposed to suggesting models. As an example of this difficulty, I am going to focus on a small and not representative portion of the excellent survey by Fehr et al [2005] on the "Neuroeconomics of Trust and Social Preferences." I found most of their arguments quite convincing, but at one point I was not convinced when they said that the evidence they present casts doubt on the claim in Samuelson [2005] that observed cooperation in the one-shot prisoners dilemma might be a consequence of players mistakenly treating the game as if it were part of a repeated interaction. This claim is an updated version of the "misperception" argument of Kenneth Binmore and Samuelson [1995], which Levine and I have previously criticized, (Fudenberg and Levine [1998], p. 98) and I do not mean to defend it here, but I do wish to question the extent to which the data cited by Fehr et al can help resolve the issue. Basically, Fehr et al note that there is more striatum activation when players cooperate with a human than when they cooperate with a computer, and also more activation than from receiving the same payment as an uncontingent reward. They argue that this shows that people are happier and derive more utility from interactions with cooperative people, and suggest that this refutes the explanation based on misperceptions. However, the Fehr et al argument seems to rest on

the assumption that whatever misperceptions Binmore and Samuelson have in mind have no effect on striatum activation, and at this point there is no reason to think that this is the case. [24]

More generally, at this point neural imaging can provide insights into the mechanisms of various behaviors and cognitive processes, and these insights may suggest useful experiments or interesting models, but we must be careful to distinguish neural correlates of a behavior from its causes. One concern is that many parts of the brain seem to be involved in processing rewards. Even in apparently simpler cases such as a monkey's valuation for various foods, where researchers have identified individual neurons whose activation is correlated with value, the mechanism by which this reward information is generated and used for decision making is not understood.[25] In particular, activations in one brain region may be the consequence of activations at other "upstream" neurons.[26] A second concern is to not confuse "biological" with "genetically determined." For example, a number of studies have shown that people in different countries and cultures tend to play differently in a range of stylized laboratory experiments, and these differences are presumably correlated with differential activation in some parts of the brain that are involved in assessing rewards and making decisions.[27] However, such a correlation would be consistent with the two cultures being genetically identical, as the differential activations could be a learned response to living in different cultures. In that case it would seem more natural to think of the activation patterns as a consequence, and not a cause, of the difference in cultures. [28]

Regardless of these issues of causality and interpretation, it is intriguing that neural imaging data can be used to predict future behavior. One of the best examples of this is in de Quervain et al [2004], who look at activations when agents decide to punish. In their experiment, players A and B are each endowed with 10 "money units." Player A can either keep his endowment of 10 or send it to player B; money sent to player B is quadrupled by the experimenter, if A sends 10, then B has 50. Next, B has the choice of sending back either nothing or half of the 50. Finally, A has the option of "punishing" B by assigning up to 20 "punishment points;" the cost to A and B of this punishment varies over treatments. In condition IC, punishment is costly to player A and costly to player B; in condition IF, punishment is free for player A and costly to B. There were 11 subjects

who punished maximally in IF. For these subjects, differences activation levels cannot be due to the chosen punishment, so it is natural to interpret them as a sign of the "reward to punishing. Strikingly, de Quervain et al [2004] find that in this pool of 11 players, activation levels (of a particular neuron in the caudate) when punishing maximally in IF are correlated with the punishments they choose in IC. That is, observing activations of different agents making the same choice in IF helps predict choice in IC. I expect that neuroeconomists will develop more of these sorts of predictions in the future. It seems too early to know just how much impact this will have on most of economics, but the potential impact is large, and the research underway is fascinating. Thus, while I don't think that every economist ought to take up neuroeconomics, I do think that anyone interested in individual choice and decision making ought to keep an eye on how it develops.[29]

Are we having fun (yet)?

It is if anything even more difficult for me to evaluate the usefulness of work on "affective forecasting" (e.g. Wilson et al [2003]) and "predictive utility" (Kahneman [1999]), which asks subjects "how happy are you now" and "how happy would you be if this outcome occurred?" This literature argues that there reports are a good measure of peoples' internal states. It also argues that people make systematic mistakes both in predicting how various outcomes will influence their happiness, and in remembering how happy they were in the past.

One possible, albeit crude and simplified, interpretation of this work goes as follows. First, reported happiness is a good measure for happiness as a subjective mental state, possibly modulo some systematic but constant differences in reporting across countries and cultures. Next, people always choose the actions that they think will make them happiest. (This view has some well-known adherents, including Daniel Gilbert.) Thus, the systematic forecasting errors found in the literature on affective forecasting show that people often make mistakes in trying to in predict how various actions will make them feel, and moreover that these mistakes lead people to take the "wrong" actions. Hence, revealed preferences are not the best guide for evaluating the effects of various government policies. More strongly, one might conclude that welfare judgments

and policy decisions should take as their objective the reported happiness of the population.

In thinking about these ideas, it is helpful to distinguish between welfare economics as political economy- how we think the government should make decisions- and welfare economics as moral philosophy- how we would advise others to behave. Even if we believe people do make systematic errors in evaluating how various choices will influence the appropriately-defined measure of their "welfare", we might not trust that the government or policy analysts would make better evaluations. For this reason, it is consistent to believe both that people make mistakes and that government policy (with a few exceptions) be based on the assumption people's actions and ex-ante predictions are the best guide to what is in their own interests. [30]

However, the situation is different when considering how we plan our own behavior or advise others how to behave, as knowledge of typical errors can help prevent one from making them. This distinction between the sorts of preferences that are considered valid in policy evaluation, and sorts of preferences that "reasonable" is important even in the absence of survey or neural data, and it holds even when people perfectly predict all of their future mental states. For example, it is consistent with the standard model to be completely impatient, and assign value 0 to all future payoffs, but as an advisor or parent I would view this extreme short-sightedness as a mistake. Indeed, a major task of parents is trying to teach children the "appropriate" weight to give to future consequences. [31] Similarly, the standard "rationality" axioms for subjective probabilities don't imply that people's probability forecasts should be calibrated in the sense of Alpert and Howard Raiffa [1982]: it is "rational" to make all of one's of 90% confidence intervals so small that they rarely contain the true value of the variable in question. Yet people whose subjective beliefs are too far from reality are deemed insane, and Bruno Biais et al [2004] find that subjects who are better calibrated (with respect to a number of questions about irrelevant statistics) do better in an experimental asset market. Thus the questions that "mistakes" pose for welfare economics are much broader than the traditional subject matter of behavioral economists.

Moving from welfare economics to the conceptually clearer task of predicting behavior, Miles Kimball and Robert Willis [2005] sketch a framework for using survey

data on happiness to better estimate and forecast consumer demand. If their project is successful, it will move survey data into the mainstream economics, but it is too early to tell if that will be the case.

REFERENCES

Ashraf, Nava; Dean Karlan, and Walter Yin. 2005. "Tying Odysseus to the Mast: Evidence from a Consumer Savings Product in the Philippines," mimeo, forthcoming in the *Quarterly Journal of Economics.*

Bajari, Patrick; H. Hong; J. Krainer, and D. Nekipelov. 2005. "Estimating Static Models of Strategic Interactions," mimeo.

Bandiera, Oriana; Iwan Barankay, and Imran Rasul. 2005. "Social Preferences and the Response to Incentives: Evidence from Personnel Data," *Quarterly Journal of Economics,* 120(3), pp. 917-962.

Barberis, Nicholas; Andrei Shleifer, and Robert Vishny. 1998. "A Model of Investor Sentiment," *Journal of Financial Economics,* 49(3), pp. 307-343.

Battigalli, Pierpaolo and Martin Dufwenberg. 2005. "Dynamic Psychological Games," mimeo.

Benabou, Roland and Jean Tirole. 2002. "Self Confidence and Personal Motivation," *Quarterly Journal of Economics,* 117(3), pp. 871-915.

Benjamin, Daniel and Jesse Shapiro. 2005. "Does Cognitive Ability Reduce Psychological Bias?", mimeo.

Bernheim, Douglas and Antonio Rangel. 2004. "Addiction and Cue-Triggered Decision Processes," *American Economic Review*, 94(5), pp. 1558-1590.

Bertrand, Marianne; Dean Karlan; Sendhil Mullinaithan; Elar Shafir, and Jonathan Zinman. 2005. "What's Psychology Worth? A Field Experiment in the Consumer Credit Card Market," mimeo.

Biais, Bruno; Dennis Hilton; Karine Mazurier, and Sébastien Pouget. 2005. "Judgmental Overconfidence, Self-Monitoring and Trading Performance in an Experimental Financial Market," *Review of Economic Studies*, 72(3), pp. 615-649.

Bjonerstedt, Jonas and Jörgen Weibull. 1995. "Nash Equilibrium and Evolution by Imitation," in *Rational Foundations of Economic Behavior*. Kenneth J. Arrow,

Enrico Colombatto, Mark Perlman and Christian Schmidt, eds. London: MacMillan.

Bodner, Ronit and Drazen Prelec. 2003. "The Diagnostic Value of Actions in a Self-Signaling Model," in *The Psychology of Economic Decisions, Vol. 1*. Isabelle Brocas and Juan D. Carillo, eds. Oxford: Oxford University Press

Camerer, Colin. 2004. "Behavioral Game Theory: Predicting Human Behavior in Strategic Situations," *Advances,* Chapter 12.

Camerer, Colin; George Lowenstein, and Drazen Prelec. 2004. "Neuroeconomics: Why Economics Need Brains," *Scandinavian Journal of Economics,* 106, pp. 555-579.

Choi, James; David Laibson, and Brigitte Madrian. 2004. "Plan Design and 401(k) Savings Outcomes," *National Tax Journal*, 57, pp. 275-298.

Crawford, Vincent. 1997. "Theory and Experiment in the Analysis of Strategic Interaction," in *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress of the Econometric Society, Volume 1*. David Kreps and Robert Willis, eds. Cambridge, England: Cambridge University.

Dekel, Eddie; Bart Lipman, and Aldo Rustichini. 1999. "Representing Preferences with a Unique Subjective State Space," *Econometrica,* 69**,** pp. 1403-1435.

Dekel, Eddie; Drew Fudenberg, and David K. Levine. 2004. "Learning to Play Bayesian Games," *Games and Economic Behavior*, 46, pp. 282-303.

Dekel, Eddie; Bart Lipman, and Aldo Rustichini. 2005. "Temptation-Driven Preferences," mimeo.

DellaVigna, Stefano and Ulrike Malmendier. 2004. "Contract Design and Self Control: Theory and Evidence," *Quarterly Journal of Economics,* 119, pp. 353-402.

DellaVigna, Stefano and Ulrike Malmendier. 2005. "Paying Not to Go to the Gym," forthcoming in the *American Economic Review.*

De Quervain, Dominique; Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schmyder, Alfred Buck, and Ernst Fehr. 2004. "The Neural Basis of Altruistic Punishment," *Science,* 305, pp. 1254-1258.

Dufwenberg, Martin and G. Kirchsteiger. 2004. "A Theory of Sequential Reciprocity," *Games and Economic Behavior,* 47, pp. 268-298.

Ellison, Glenn. 2005. "Bounded Rationality in Industrial Organization," mimeo.

Ellison, Glenn and Sarah F. Ellison. 2005. "Search, Obfuscation, and Price Elasticities on the Internet," mimeo.

Ellison, Glenn and Drew Fudenberg. 1993. "Rules of Thumb for Social Learning," *Journal of Political Economy*, 101, pp. 612-643.

Ellison, Glenn and Drew Fudenberg. 1995. "Word of Mouth Communication and Social Learning," *Quarterly Journal of Economics*, 110, pp. 93-126.

Eyster, Eric and Matthew Rabin. 2005. "Cursed Equilibrium," mimeo, forthcoming in *Econometrica.*

Fehr, Ernst; Urs Fischbacher, and Michael Kosfeld. 2005. "Neuroeconomic Foundations of Trust and Social Preferences," mimeo, forthcoming in the *American Economic Review.*

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics,* 114(3), pp. 817-868.

Frederick, Shane; George Loewenstein, and Ted O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review," *Journal of Economic Literature,* reprinted in *Advances.*

Fudenberg, Drew and David Kreps. 1993. "Learning Mixed Equilibria," *Games and Economic Behavior*, 5, pp. 320-367.

Fudenberg, Drew; David Kreps and David K. Levine. 1988. "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory*, 44, pp. 354-380.

Fudenberg, Drew and David K. Levine. 1983. "Subgame-Perfect Equilibria of Finite- and Infinite-Horizon Games," *Journal of Economic Theory*, 31, pp. 251-258.

Fudenberg, Drew and David K. Levine. 1998. *Theory of Learning in Games.* Cambridge, MA: MIT Press.

Fudenberg, Drew and David K. Levine. 1998. "Learning in Games: Where Do We Stand?" *European Economic Review*, 42, pp. 631-639.

Fudenberg, Drew and David K. Levine. 2005. "A Dual-Self Model of Impulse Control," mimeo.

Gabaix, Xavier and David Laibson. 2005. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets," mimeo.

Glaeser, Edward L. 2005. "The Political Economy of Hatred," *Quarterly Journal of Economics,* 120(1), pp. 45-86.

Gul, Faruk and Wolfgang Pesendorfer. 2001. "Temptation and Self Control," *Econometrica*, 69, pp. 1403-1436.

Haruno, Masahiko; Tomoe Kuroda, Kenji Doya, Keisuke Toyama, Minoru Kimura, Kazuyuki Samejima, Hiroshi Imamizu, and Mitsuo Kawato. 2004. "A Neural Correlate of Reward-Based Behavioral Learning in Caudate Nucleus: A Functional Magnetic Resonance Imaging Study of a Stochastic Decision Task," *The Journal of Neuroscience*, 24, pp. 1660-1665.

Hossain, Tamjir and John Morgan. 2005. "Plus Shipping and Handling: Revenue (Non)-Equivalence in Field Experiments on eBay," forthcoming in *Advances in Economic Analysis and Policy.*

Kagel, John H. 1995. "Auctions," in *The Handbook of Experimental Economics*. Alvin Roth and John Kagel, eds. Princeton: Princeton University Press.

Kagel, John H. and Jean-Francois Richard. 2001. "Super-Experienced Bidders in First-Price Common-Value Auctions: Rules of Thumb, Nash Equilibrium Bidding, and the Winner's Curse," *The Review of Economics and Statistics,* 83(3), pp. 408-419.

Kahneman, Daniel. 1999. "Objective Happiness," in *Well-Being: The Foundations of Hedonic Psychology*. Daniel Kahneman, Ed Diener, Norbert Schwarz, and Daniel Kahnemann, eds. New York: Russell Sage Foundation.

Kamenica, Emir. 2005. "Contextual Inference in Markets: On the Informational Content of Product Lines," mimeo.

Kimball, Miles and Robert Willis. 2005. "Utility and Happiness," mimeo.

Koszegi, Botond and Matthew Rabin. 2004. "A Model of Reference-Dependent Preferences," mimeo.

Kreps, David. 1990. *Game Theory and Economic Modeling*. Oxford: Oxford University Press.

Kuksov, Dmitri and J. Miguel Villas-Boas. 2005. "When More Alternatives Lead to Less Choice," mimeo.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics,* 112, pp. 443-477.

Laibson, David. 2001. "A Cue-Theory of Consumption," *Quarterly Journal of Economics,* 116, pp. 81-120.

Laibson, David and Leeat Yariv. 2005. "Safety in Markets: An Impossibility Theorem for Dutch Books," mimeo.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics,* 1, pp. 593-622.

Lowenstein, George. 1996. "Out of Control: Visceral Influences on Behavior," *Organizational Behavior and Human Decision Process,* 65, pp. 272-292.

Lowenstein, George and Ted O'Donoghue. 2004. "Affective and Deliberative Processes in Economic Behavior," mimeo,

Machina, Mark. 1984. "Temporal Risk and the Nature of Induced Preferences," *Journal of Economic Theory,* 33, pp. 199-231.

McClure, Samuel M.; David Laibson; George Loewenstein, and Jonathan D. Cohen. 2004. "Separate Neural Systems Value Immediate and Delayed Monetary Rewards," *Science*, 306, pp. 503-507.

McIntosh, D. 1969. *Foundations of Human Society*. Chicago: University of Chicago Press.

McKelvey, Richard and Thomas Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, 10(1), pp. 6–38.

Mischel, Walter; Yuichi Shoda, and Monica L. Rodriguez. 1989. "Delay of Gratification in Children," *Science,* 244, pp. 933-938.

Mullainathan, Sendhil and Andrei Shleifer. 2004. "The Market for News," mimeo.

Muraven, Mark; Dianne M. Tice, and Roy F. Baumeister. 1998. "Self-Control as a Limited Resource," *Journal of Personality and Social Psychology,* 74(3), pp. 774-789.

Muraven, Mark and Roy F. Baumeister. 2000. "Self-Regulation and Depletion of Limited Resources: Does Self-Control Resemble a Muscle?" *Psychological Bulletin,* 126**,** pp. 227-259.

O'Doherty, John. 2004. "Reward Representations and Reward-Related Learning in the Human Brain: Insights from Neuroimaging," *Current Opinion in Neurobiology,* 14, pp. 769-776.

O'Donoghue, Ted and Matthew Rabin. 1999. "Doing It Now or Later," *American Economic Review*, 89, pp. 103-124.

O'Donoghue, Ted and Matthew Rabin. 2001. "Choice and Procrastination," *Quarterly Journal of Economics*, pp. 121-160.

O'Donoghue, Ted and Matthew Rabin. 2003. "Optimal Sin Taxes," mimeo.

Padoa-Schioppa, Camillo and John Assad. 2005. "Neuronal Processing of Economic Value in OrbitoFrontal Cortex," mimeo.

Pakes, Ariel; Michael Ostrovsky, and Steve Berry. 2005. "Simple Estimators for the Parameters of Discrete Dynamic Games," mimeo.

Platt, Michael L. and Paul Glimcher. 1999. "Neural Correlates of Decision Variables in Parietal Cortex," *Nature,* 400, pp. 233-238.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics," *American Economic Review,* 83, pp. 1281-1302.

Rabin, Matthew. 1997. "Fairness in Repeated Games," mimeo.

Rabin, Matthew. 2000. "Risk Aversion and Expected-Utility Theory: A Calibration Theorem," *Econometrica*, 68(5), pp. 1281-1292.

Rabin, Matthew. 2002. "The Law of Small Numbers," *Quarterly Journal of Economics,* 117, pp. 775-816.

Rabin, Matthew and Joel C. Schrag. 1999. "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics,* 114, pp. 37-82.

Samuelson, Larry. 1998. *Evolutionary Games and Equilibrium Selection.* Cambridge, MA: MIT Press.

Samuelson, Larry. 2005. "Foundations of Human Society: A Review Article," forthcoming in the *Journal of Economic Literature.*

Schlag, Karl. 1998. "Why Imitate, and If So, How? A Boundedly Rational Approach to Multi-Armed Bandits," *Journal of Economic Theory,* 78, pp. 130-156.

Schlag, Karl. 2005. "Competing for Boundedly Rational Consumes," mimeo.

Schulz, Wolfram; Peter Dayan, and P. Read Montague. 1997. "A Neural Substrate of Prediction and Reward," *Science*, 275, pp. 1593–1599.

Shafir, Eldar; Peter Diamond, and Amos Tversky. 1997. "Money Illusion," *Quarterly Journal of Economics,* reprinted in *Advances,* chapter 17.

Shapiro, Jesse. 2005. "Fooling Some of the People Some of the Time," mimeo.

Shefrin, Hersh and Richard Thaler. 1988. "The Behavioral Life-Cycle Hypothesis," *Economic Inquiry,* 26, pp. 609-643

Shiv, Baba and Alexander Fedorikhin. 1999. "Heart and Mind in Conflict: The Interplay of Affect and Cognition in Consumer Decision Making," *Journal of Consumer Research*, 26(3), pp. 278-292.

Singer, Tania; Stefan J. Kiebel, Joel S. Winton, Raymond J. Dolan, and Chris D. Frith. 2004. "Brain Responses to the Acquired Moral Status of Faces," *Neuron,* 41(4), pp. 653-662.

Smallwood, Dennis and John Conlisk. 1979. "Product Quality in Markets Where Consumers are Imperfectly Informed," *Quarterly Journal of Economics*, 93(1), pp. 1-23.

Spence, Michael and Richard Zeckhauser. 1972. "The Effect of the Timing of Consumption Decisions and the Resolution of Lotteries on the Choice of Lotteries," *Econometrica*, 40(2), pp. 401-403.

Spiegler, Rani. 2005. "Competition Over Agents with Boundedly Rational Expectations," mimeo.

Stigler, George. 1965. "The Development of Utility Theory," in *Essays in the History of Economics*. Chicago: Chicago University Press.

Strotz, Robert H. 1955. "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies,* 23(3)**,** pp. 165-180.

Thaler, Richard and Hersh Shefrin. 1981. "An Economic Theory of Self-Control," *Journal of Political Economy,* 89(2), pp. 392-406.

Tversky, Amos and Daniel Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases," *Science,* 185, pp. 1124-1131.

Ward, Andrew and Traci Mann. 2000. "Don't Mind If I Do: Disinhibited Eating Under Cognitive Load," *Journal of Personality and Social Psychology,* 78, pp. 753-763.

Weibull, Jörgen. 1995. *Evolutionary Game Theory.* Cambridge, MA: MIT Press.

Wertenbroch, Klaus. 1998. "Consumption Self-Control via Purchase Quantity Rationing of Virtue and Vice," *Marketing Science*, 17, pp. 317-337.

Wilson, Timothy, David Myers, and Daniel Gilbert. 2003. "How Happy Was I, Anyway: A Retrospective Impact Bias," *Social Cognition*, 21, pp. 407-432.

Yariv, Leeat. 2005. "I'll See It When I Believe It: A Simple Model of Cognitive Consistency," mimeo.

Young, Peyton. 2004. *Strategic Learning Theory*. Oxford: Oxford University Press.

Crawford, Vincent, and Nagore Iriberri. 2005. "Level-k Auctions: Can a Non-Equilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" Mimeo, http:weber.ucsd.edu/~vcrawfor/#MostRecent

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir . 1991. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," American Economic Review, 81, pp. 1068-1095.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, (Eds.) Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. Oxford University Press, 2004 .

[1] I thank Daniel Benjamin, Colin Camerer, Peter Diamond, Glenn Ellison, Ernst Fehr, Faruk Gul, Emir Kamenica, David Laibson, Ariel Pakes, Parag Pathak, Wolfgang Pesendorfer, Drazen Prelec, Matt Rabin, and Jesse Shapiro for helpful comments and conversations. I am particularly grateful to David K. Levine for years of collaboration; this essay draws heavily on our joint work. NSF grant SES-04-26199 provided financial support.

[2] Department of Economics, Harvard University

[3] The research articles include three from the *American Economic Review,* one from the *Journal of Political Economy,* and nine from the *Quarterly Journal of Economics,* which reflects the *QJE's* early commitment to the field.

[4] Prospect theory is handicapped by the lack of an accepted model of how reference points are determined; ongoing work of Botond Koszegi and Rabin [2004] tries to provide one.

[5] The quasi-hyperbolic model suffers from spurious equilibria. Drew Fudenberg and David Levine [2005] argue that an alternative "dual-self" model can account for the regularities usually explained with quasi-hyperbolic discounting and additional regularities as well, while generating unique predictions in standard economic models.

[6] The Fehr and Schmidt model derives its simplicity from the restriction that the way agents evaluate a given social outcome is independent of the other outcomes that could have occurred. On the other hand, unlike most models in this area, Fehr and Schmidt do explicitly allow the agents to differ in the magnitude of their concern for others. Of course even classic rational agents can have heterogeneous utility functions, but heterogeneity is even more important in settings where the classical assumptions are violated: For example, there is only one way to do Bayes rule correctly, but many ways to do it wrong. Similarly, in most experiments there is only one outcome that maximizes a subject's money payoff, but there could be many outcomes that maximized a function that depends on the money payoffs of all players.
Matthew Rabin [1993] presents a static model of fairness, where play depends in part on the anticipated 'friendliness' of the opponent's action; this concern cannot be captured in the Fehr and Schmidt model because the friendliness of one action depends on the payoffs to the alternatives. Rabin 1997 and Dufwenberg and George Kirchsteiger [2004] extend the static model to extensive- form games while maintaining the assumption that social preferences are common knowledge Given the range of possible social preferences, the assumption that the social preferences are common knowledge seems too strong; Levine [1998] and Pierpaolo Battigalli and Dufwenberg [2005] explore models that relax this.

[7] This is called "choice overload" in the behavioral literature. It can occur even with standard rational agents. true even without accounting for cognitive limitations on the part of economists, as shown by e.g. Dmitri Kuksov and Miguel Villas-Boas [2005] and Emir Kamenica [2005], so my argument does not rely on the assumption that economists are boundedly rational.

[8] Game theory is more of a methodology than a body of empirical facts, but its history in some ways can illustrate the evolution I have in mind. When I was a graduate student in the late 1970's, the only game theory taught in first-year graduate classes at Harvard and MIT was Cournot equilibrium, which was fit uneasily into the theory sequence to show how prices approach the competitive limit as the number of firms grows. The reason that game theory has expanded its share of first-year classes is that its has evolved from a collection of examples and special cases to a much more general framework Early authors such as Cournot, Bertrand, Hotelling and Stackelberg offered stand-alone solution concepts for particular situations, but the field did not really exist before the development and analysis of general constant-sum games by John von Neumann, and it did not become useful for economic applications until John Nash formulated the general Nash equilibrium concept, which made it possible to see the solutions proposed by Cournot, Bertrand, and Hotelling as special cases corresponding to different strategy spaces. But the Nash concept on its own proved too limited to provide a foundation for analyzing many important situations, which may be why game theory had not made much of an inroad into economics by 1970. Economists in the 1960's and 70's extended the application of game theory to the study of commitment and timing on an ad-hoc basis by making assertions about which out-of-equilibrium responses players should view as credible; subsequent authors grounded this analysis using refinements of Nash equilibrium such as subgame- perfect equilibrium. Over the same period, economists moved from ad-hoc analyses of games of incomplete information to analyses based on Harsanyi's reformulation; game theory became a major part of graduate classes once the combined power of the Nash, Harsanyi and Selten work became evident.

[9] Most direct evidence of framing effects has so far come from laboratory settings, but the Bertrand et al [2005] study of ads for bank loans shows that framing can have significant economic effects in the field.
[10] This is consistent with evidence (such as Pavlov's bell) that the impulsive short-run self responds to learned behavioral cues in addition to direct stimuli. Modern physiological research is making progress in identifying some of the brain chemistry that reflects the response to these stimuli, see, for example, Haruno et al [2004]. Camerer, Lowenstein and Prelec [2004] say that "roughly speaking, it appears that similar brain circuitry (dopaminergic neurons in the midbrain) is active for a wide variety of rewarding experiences (including) money rewards."
[11] This is true for example of the many studies cited in Chapter 6, (Shane Frederick, George Loewenstein, and Ted O'Donoghue [2004]) p. 173..
[12] Rabin [2000] calls this "the paradox of risk aversion in the small and in the large." Fudenberg and Levine [2005] propose an explanation of this paradox based on the ideas of self-control costs and a cash-in-advance constraint on purchases.
[13] McLure et al [2004] argue that their subjects treat Amazon.com gift certificates as if they represented immediate consumption. The Shiv and Ferorihkin [1999] experiment suggests that a ticket for chocolate cake (to be consumed in a few minutes) is more tempting when accompanied by the display of a chocolate cake than when it is accompanied only by a photograph of a cake.
[14] The idea that not all play is equilibrium play has a long history in the game theory literature, and the papers cited in the section on game theory (Chapters 12 and 13) make very little use of observations or insights from psychology (although Chapter 13 does discuss social preferences.) Thus, while it is better economics to acknowledge the likelihood of non-equilibrium play in some settings, it is not clear why this falls under the definition of behavioral economics as "modifying […] standard assumptions in the direction of greater *psychological* realism." (emphasis added).
[15] The books by Fudenberg and Levine [1998], Kreps [1990], Larry Samuelson, [1998], Jörgen Weibull [1995] and Peyton Young [2004] discuss theoretical models of learning and other sorts of non-equilibrium adjustment procedures. Many chapters in the Handbook of Experimental Economics (edited by John Kagel and Alvin Roth, 1995) discuss the effect of learning on play in game theory experiments, as do Chapters 12 . ry experiments.
[16] Glenn Ellison and Fudenberg [1993], [1995] discuss models of boundedly-rational social learning about the state of Nature; the models of Bjornstedt and Weibull [1995] and Schlag [1998] can be viewed as a form of boundedly-rational social learning about opponents' strategies
[17] Work to date on estimating payoff functions using the equilibrium assumption is only partially encouraging in this regard. Bajari et al [2005] provide conditions under which payoff functions can be estimated without parametric restrictions. They require that the privately observed shocks to payoffs are an additive term that depend only on the player's own action and independent over time and across players. as in Fudenberg and Kreps [1993] and Richard McKelvey and Thomas Palfrey [1995], and in addition that the distribution of the shocks is known up to a parameter, and that the game has a unique equilibrium. Ariel Pakes, Michael Ostrovsky, and Steve Berry [2005]show how to estimate the Markov- perfect equilibrium for dynamic games of entry and exit; their procedure relies on functional forms for the payoff functions and also the assumption that play is Markov with respect to the designated state space. It isn't clear at this point how to distinguish non-equilibrium play from play of an equilibrium that isn't Markov, or equivalently is Markov on a different state space. Thus, while progress is being made in estimating equilibria, it may be difficult to find conditions that will equilibrium from non-equilibrium play in field data.
[18] Dekel et al [2004] provide a learning-theoretic analysis of how the outcomes of Bayesian games depends on what agents observe when the game is played. Their analysis assumes that players' equilibrium beliefs are consistent with what they observe when the equilibrium is played; one possible way to provide foundations for the equilibrium concepts I criticize here would be to modify the Dekel et al setup to reflect a specified form of cognitive error.
[19] John Kagel [1995] surveys many results on how the extent of the winner's curse varies as the subjects gain experience; Kagel and Jean-Francois Richard [2001] argue that "super-experienced" bidders are not subject to this curse. If it turned out that a wide range of data could be explained with a fixed value of $\psi < 1$, there would be a positivist case for using cursed equilibrium despite its lack of foundations, but in practice the extent of cursedness varies. Recently Vincent Crawford and Nagore Iriberri [2005] have

proposed an alternative explanation for overbidding by inexperienced subjects in common-value auctions that is based on "k level reasoning" instead of equilibrium analysis.

[20] Intuitively, when consumption must be set ahead of time, agents prefer lotteries with less variance, so neither the independence axiom nor the substitution axiom (i.e., the reduction of compound lotteries to simple ones) should be expected to hold; see Spence and Zeckhauser [1972] and Machina [1984] for details. Gul and Pesendorfer do not specify any particular processes that might underlie their assumptions on preferences, so in particular they do not assume that self-control requires effort; the argument above shows a difficulty with one possible interpretation.

[21] See Ellison [2005] for a recent survey of the bounded rationality literature.

[22] Note that a significant fraction of consumers do seem to consider gasoline mileage when buying cars. This is probably related to the fact that mainstream newspapers report on gas prices, and as do their car reviews, while printer reviews and discussions of ink prices are found only in more specialized media, but the causality of the relationship is unclear.

[23] This conclusion rests on the belief that models with more accurate assumptions make more accurate predictions, so that the "'second-best problem" does not arise. As a personal example, the neuro-imaging work of McLure et al [2004] was the impetus for my work with Levine on a dual-self model of impulse control.

[24] To expand on these points, suppose that we look at activations in the first period in which a player chooses to defect. If striatum activation is different here than in previous periods, the activation is not simply coding the pleasure of interacting with cooperators. As a second thought experiment, suppose we pair human subjects with human stooges who always play C. Some of these subjects will learn that it pays to always play D. If, as I expect, the striatum activation will be less than that of someone who always plays C, it will be hard to interpret the activation as the hedonic utility of meeting a friendly opponent. The interpretation of data on striatum activation is further clouded by the fact that reward-related activations are greater for unexpected than for expected rewards (Schulz et al [1997]). It is true that Singer et al [2004] show that viewing faces of people who previously cooperated activates a number of other reward-related areas, but once again this fact seems open to several interpretations.

[25] See Camilo Padoa-Schioppa and John Assad [2005].

[26] See Platt and Paul Glimcher [1999] and John O'Doherty [2004].

[27] Roth, Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir [1991] was the first formal study of cross-country differences in game theory experimemnts; the book by Joseph Henrich, Robert Boyd, Samuel Bowles, Camerer, Fehr, and Herbert Gintis [2004] surveys a number of more recent studies.

[28] Of course it is also possible that over a long enough time span the difference in cultures could have an effect on genetics, in which case the causality would be more complicated

[29] The interaction between economic theory and neuroscience is not one-way, as decision and game-theoretic models have been used to predict how the firing of certain neurons varies with changes in payoffs, as for example in Platt and Glimcher [1999].

[30] Recent work by James Choi et al [2004] on the impact of default options and O'Donoghue and Rabin [2003] on "conservative paternalism" explore how to set policies that let help correct errors of "irrational" agents with minimal interference to the choices of "rational" ones.

[31] The primary reference for this assertion is folk wisdom and personal experience, but it is interesting to note that it accords with the findings of Mischel, Shoda and Rodriguez [1989] that "those 4-year-old children who delayed gratification longer in certain laboratory situations developed into more cognitively and socially competent adolescents." Benjamin and Shapiro [2005] find that students who do better on standardized math tests are more patient over short-term trade-offs and less risk-averse over small gambles. The causal relationship between patience and various measures of performance remains an open question, see Benjamin and Shapiro for some thoughts on this issue.