View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by UCL Discovery

Introduction

Cleaning Test Tubes

When I started doing experimental work in the 1980s, the subject was in its infancy among economists, but one set of findings was thought to be rock solid. Game theory doesn't work in the laboratory. People don't play Nash equilibria. They don't use their maximin strategies in two-person, zero-sum games. They even cooperate in the Prisoners' Dilemma.

But the rock on which these certitudes were based has crumbled away. It is true that unmotivated subjects in unfamiliar situations don't play as game theory predicts. So if game theory had to predict interactive human behavior under all circumstances to be worthy of attention, it would indeed be a failure. But who would want to claim of any theory that it work in all environments? Just as Newton's laws of motion don't predict well at the bottom of the sea, so game theory can't reasonably be expected to work in environments in which its tacit assumptions have no chance of being true. So what is the kind of environment in which we might reasonably expect game theory to predict well?

Favorable Environments

A conservative specification of a favorable experimental environment for game theory requires that all three of the following criteria be satisfied:

- The game is simple, and presented to the subjects in a user-friendly manner.
- The subjects are paid adequately for performing well.
- Sufficient time is available for trial-and-error learning.

Critics rightly say that these criteria are too stringent to cover all the economic situations to which game theory gets applied, but who would want

to defend each and every crazy application of the theory? Such enthusiasts certainly exist, but they seem to me no less misguided than the skeptics who determinedly turn a blind eye to any evidence that isn't hostile to game theory.

My three environmental criteria aren't intended to be hard-and-fast necessary and sufficient conditions for game theory to predict human behavior. Game theory sometimes works when one or more of the criteria aren't satisfied. It sometimes fails when all three criteria are satisfied. However, the successes are now so well established that the first response to finding that a game-theoretic prediction fails in a laboratory when all three criteria hold is to ask the same question that chemists ask if something unexpected happens when they mix reagents together:

Did I clean my test tubes properly?

Bargaining

My own attempts to work with clean test tubes in the laboratory largely fall into two categories: experiments on bargaining and experiments on auctions. The latter work was all conducted on behalf of governments and commercial enterprises. I don't report on it here, partly for reasons of confidentiality, but mostly because nobody seems to doubt that game theory is a useful guide to predicting human bidding behavior. All but one of the papers from my experimental repetoire that make up this volume are therefore devoted to tests of game-theoretic models of bargaining.

The case of bargaining is a particularly challenging case for game theory—perhaps the most challenging case of all. Everyone agrees that human behavior in real-life bargaining situations is governed at least partly by fairness considerations that we don't understand very well. But what happens when such fairness considerations conflict with game-theoretic predictions in the laboratory? Will people adapt their behavior so that they end up playing a novel bargaining game strategically? Or must we expect them simply to play fair?

Even when the test tubes are clean, experiments on bargaining models therefore come with the dice loaded against game theory. But I hope that the evidence to be presented will justify my boldness in defending the theory in a case where skeptics think the arguments in its favor are at their weakest.

The Behavioral Challenge

I think the claims made for game theory in the previous section would be uncontroversial if the issues weren't clouded by an emotional debate that seems to me entirely orthogonal to the issue of whether or not game theory works. This is the question of whether people are inherently selfish, or whether they care about those around them.

Although I think the question isn't central to the issue of whether game theory works, it isn't possible to get a hearing nowadays for the kind of experimental results I report here without confronting this controversy, since the behavioral economists who emphasize the importance of other-regarding or social preferences commonly believe that their findings represent a threat to traditional game theory.

No amount of denial seems capable of altering their conviction that game theorists like myself must necessarily believe that human beings have no interest whatever in playing fair when the chips are down. I sometimes try to shake their certitude by pointing out that I have probably written more on how and why fairness matters than any economist ever, but I find this gets me nowhere because the reasons why I think social preferences matter are so different from theirs (Binmore 1994, 1998, 2005).

The rest of this introduction is therefore devoted to making three points. The first is that the behavioral school could well be right in claiming that people have strong other-regarding preferences without their results presenting any challenge to game theory at all. The second is that one can believe that social preferences matter enormously in human conduct without agreeing at all with the behavioral school about how they matter. The third is that the level of scientific rigor thought adequate by some leading proponents of the behavioral school represents no improvement on that of the experts who used to claim that people nearly always cooperate in the Prisoners' Dilemma.

Are People Selfish?

Should we model the people who enter our laboratories as seeking to maximize the money in their own pockets? Or should we model them as maximizing a more complicated utility function, whose arguments take account of the welfare of others?

I think one might as well ask when you stopped beating your wife. In discussing the behavior of inexperienced laboratory subjects, the first question isn't what kind of utility function they are maximizing, but

whether they can sensibly be seen as maximizing anything at all (Gigerenzer 2004).

The behavior of laboratory subjects often changes markedly over time as they learn the ropes in a new experiment. We can make the maximizing hypothesis into a tautology by introducing utility functions that correspondingly change with time, but who thinks that this would be a worthwhile activity? It is true that abandoning the maximizing hypothesis implies that we have to look beyond traditional economic theory for explanations of how inexperienced subjects learn to play games, but I see no reason why we should imagine that psychology and sociology are irrelevant when trying to make sense of boundedly rational behavior.

Only after the learning phase is over can we expect to find subjects at a Nash equilibrium, each behaving as though trying to maximize his or her own utility function given the behavior of the other subjects. But do we then not find them simply maximizing money?

The answer is that this is indeed what we usually do observe—provided that the monetary payoffs are chosen to be sufficiently large. However, we can't deduce that real people therefore don't have other-regarding preferences, because part of the reason that experimenters like myself believe that the monetary payoffs need to be relatively large is to swamp whatever other-regarding preferences may be present (Vernon Smith 1976).

The school of behavioral economists who insist that other-regarding preferences matter in real life therefore have nothing to fear from experiments that show that game theory often works—unless they want to claim that subjects care so enormously about other people that it is always impossible to control their preferences in the laboratory by paying relatively large sums of money. They therefore don't need to seek to discredit game theory by endlessly drawing attention to the fact that it mostly doesn't work for inexperienced and underpaid subjects.

Nor have game theorists anything to gain from denying that the payoffs in real-life games might sometimes be derived from other-regarding preferences. Game theory is the same whether it is used to advise Saint Francis of Assisi or Attila the Hun. We simply recognize the difference between Attila and Saint Francis by writing different payoffs in the games we model them as playing.

Prisoners' Dilemma

The Prisoners' Dilemma is the most famous of all the toy games that game theorists use to illustrate their ideas. In the payoff table of figure 1,

	dove	*hawk*
dove	2 2	3* 0
hawk	0 3*	1* 1*

Figure 1
Prisoners' Dilemma

Adam's payoffs are in the bottom left of each cell and Eve's are in the top right. Adam chooses a row and Eve chooses a column. Each then receives the payoff in the cell their choices jointly determine.

The starred payoffs indicate best replies. Thus, if Eve chooses *dove*, Adam can get a payoff of 1 by choosing *dove*, and a payoff of 3 by choosing *hawk*. Since $3 > 1$, Adam's payoff of 3 is starred to show that *hawk* is his best reply to Eve's choice of *dove*. Both payoffs are starred in the cell that arises when both players choose *hawk*, which implies that the strategy pair (*hawk*, *hawk*) is a Nash equilibrium, since each player is then making a best reply to the strategy choice of the other.

The idea that it is rational to play *hawk* in the Prisoners' Dilemma has historically generated great hostility, since everyone can see that both players would get more if both played *dove*. All kinds of fallacies have therefore been invented in hopeless attempts to prove that it can be rational to play something other than the Nash equilibrium of the game (Binmore 1994). Fortunately, this activity seems to have gone out of fashion for the moment, but it remains popular to claim that laboratory experiments show that the game-theoretic analysis of the Prisoners' Dilemma has no practical relevance.

If this is your aim, then it is very easy to organize an experiment that meets your requirements. Just as alchemists can "refute" the predictions of modern chemistry by mixing their reagents in dirty test tubes, so one can "refute" game theory by confusing the subjects with complicated instructions, or by providing them with inadequate incentives, or with too little time to get to grips with the problem that has been set.

One response to such criticism is that our test tubes need to be dirty, because that's how they are in real life. Those of us who clean our metaphorical test tubes can then be accused of "fixing" our experiments to get the results we want. But who would apply the same reasoning to chemistry experiments?

Incentives

A much-quoted experiment of Robert Frank illustrates the genre I am criticizing. Despite what is commonly said, even inexperienced subjects cooperate only about half the time in the one-shot Prisoners' Dilemma (Camerer 2003, p. 46).[1] However, in Frank's (2004) modification of the usual experimental design, subjects were allowed to fraternize for half an hour before playing. It turned out that relatively few subjects were then willing to cheat on their partners by playing *hawk* after promising to play *dove*, although they could gain a dollar by doing so.

But of course not! Who is going to metaphorically stab even a new friend in the back for one measly dollar? Even Attila the Hun wouldn't bother.

Sometimes such experiments are defended with the claim that it doesn't matter whether or not you pay the subjects, as the results turn out much the same either way. Such apologists can point to experiments in which behavioral "anomalies" remain unaffected as the rewards get large. In the Ultimatum Game they can get very large indeed (Cameron 1999).

But the fact that the size of the reward is irrelevant in some environments doesn't imply that it is irrelevant in most environments. Right at the beginning of modern experimental economics, Vernon Smith (1976) observed that the amount subjects are paid can make a substantial difference in economic experiments. If this weren't true most of the time, economists presumably would have learned by now that they don't need to spend large sums of their hard-to-get research money incentifying their experimental subjects.

My own most striking experience was when I ran laboratory experiments to test a design for a major British telecom auction for which I was responsible (which eventually raised $35 billion). The pilot experiments came nowhere near the efficient outcome predicted by game theory, but when we doubled the financial incentives—so that subjects went home with about $60 on average rather than $30—the results were suddenly very close to the theoretical predictions.

Experience

Incentives therefore matter much of the time, but what I think matters most is experience. Here again, Vernon Smith (1991) was early on the scene. In a classic experiment, he found that subjects needed to be

recalled to the laboratory for three separate sessions of experience with an artificial financial market before they finally learned not to create bubbles.

Despite what is commonly said to the contrary by those who don't know or care about the literature, the case of the Prisoners' Dilemma and other toy games that can be thought of as modeling the private provision of public goods is particularly clear.[2] The huge number of experimental studies available in 1995 was surveyed both by John Ledyard (1995) and by David Sally (1995), the former for Roth and Kagel's authoritative *Handbook of Experimental Economics*. Camerer's (2003, p. 46) more recent *Behavioral Game Theory* endorses their conclusions.

It is true that inexperienced subjects often cooperate (by playing *dove*), but as the subjects gain experience, they defect more and more (by playing *hawk*), until about 90 percent are defecting. One can disrupt the march toward equilibrium by intervening in various ways, but when active intervention ceases, the march resumes.

Figure 2 is from a paper by Fehr and Gächter (2000). It is included to emphasize that these conclusions are uncontested even by authors who are commonly quoted with a view to discrediting traditional game theory. The first ten periods show the standard decline in the average contribution as the subjects gain experience in a regular public goods game.[3] In the final round nearly everyone contributes nothing.

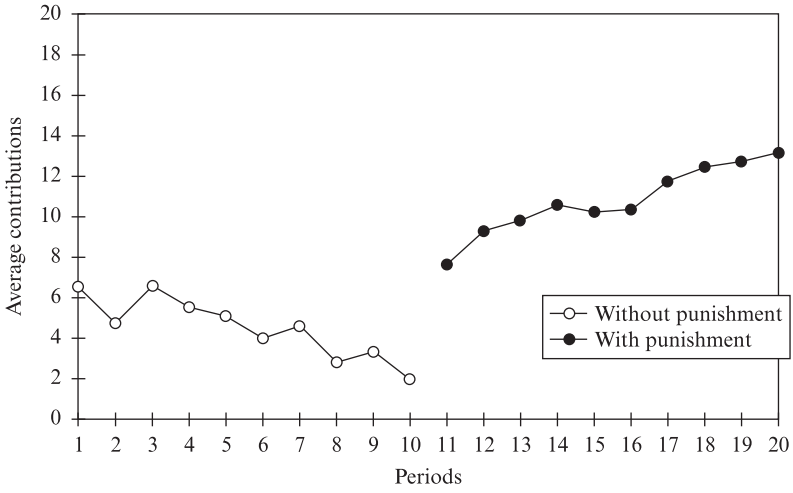

Figure 2
Public goods experiments before and after punishment (Fehr and Gächter 2000a, fig. 3B).

What Does Game Theory Predict?

But what about the behavior in the second ten periods of Fehr and Gächter's (2000) experiment?

In this part of the experiment the game is changed so that the subjects can pay a relatively small amount to reduce the payoff of free riders by a relatively large amount. They wouldn't take advantage of this opportunity to punish free riders in a subgame-perfect equilibrium of the one-shot game, but the data from the second ten periods of the experiment show that on the contrary, the threat of punishment induces the subjects to contribute more and more as they gain experience of the new game.

Behavioral economists take such data as proof that people have other-regarding preferences, but it isn't hard to think of other reasons why the equilibrium that behavioralists identify as the orthodox prediction isn't appropriate. For example, there isn't any particular reason why an adjustment process should converge on the subgame-perfect equilibrium of a one-shot game when other Nash equilibria are available—which they usually are (appendix C at the end of this volume). Nor is it obvious that we should be looking at Nash equilibria of the one-shot game when small groups of subjects play repeatedly (chapter 8).

Even if one insists on looking only at subgame-perfect equilibria of the one-shot game, it is unnecessary to postulate more than a small other-regarding component in the subjects' utility functions to create a game with a cooperative equilibrium. For example, Jakub Steiner (1972) offers a model in which the subjects feel just a little angry with free riders. He then describes an equilibrium in which only the worst free rider would get punished. The small cost of punishing then becomes tiny because it is shared among all the punishers. But the punishment is enough to support an equilibrium without free riding in the one-shot game, since a player who is the only free rider will necessarily be the most guilty (chapter 8).

No Convergence

However, the reason for spending time on the second ten periods of Fehr and Gächter's experiment isn't so much to question their claims about what game theory ought to predict about the equilibrium on which their subjects might eventually converge if the game were repeated often enough. It is to point out that although the subjects' behavior converges fairly well to the standard result in the experiment of the first ten periods, their behavior in the experiment of the second ten periods hasn't got close to converging on anything at all.

The graph of figure 2 shows the subjects' average behavior changing fairly rapidly over time. Nor is there any sign of the subjects coalescing around the average. As the authors point out, the distribution of contributions in the final round is spread out over the whole range of possibilities. It is therefore premature to ask to what extent the subjects should be seen as revealing other-regarding or selfish utilities in the second experiment. The subjects' behavior isn't consistent with maximizing any time-independent utility function at all.

This comment may seem too obvious to be worth making, but it isn't at all popular. Neoclassical economists are often as impatient as behavioral economists with the idea that people need time to adapt to a new game because they think of learning as an exclusively intellectual activity—and what is there to learn in such a simple game?

But I think the kind of learning that is going on is more akin to a sailor's learning not to walk with a rolling gait when he comes ashore after a long voyage. His mind knows perfectly well that he is on dry land, but his body hasn't figured out yet that this implies that he doesn't need to keep making ready for the next wave.

Coming Ashore

Everyone agrees that much of our interaction with other human beings is governed by *social norms*. I see such norms as analogues in social life of a sailor's rolling gait.

Just as a sailor's rolling gait is an efficient adaptation to the need to be ready for the next wave during a long voyage, so game theorists of my persuasion think it likely that cultural evolution has shaped our social norms so that their use mostly results in our coordinating on efficient equilibria in the real-life games that we play every day with those around us.

Of course, we are seldom any more aware that this is what we are doing than a sailor is conscious of walking oddly. We usually aren't even conscious that we are playing a game. For ordinary human beings, using a social norm is a piece of habituated behavior that is triggered by appropriate environmental cues.

Habits are hard to shake off—especially if you are unconscious that you have a habit in the first place. So when the *framing* of an experiment triggers the appropriate environmental cues, we often respond with the habituated response: no matter how ill-adapted it may be to the actual game being played in the laboratory. Like a sailor stepping ashore, we

still roll with the waves, even though there are no longer any waves with which to roll.

I therefore think that Kahneman and Tversky's (1988) emphasis on the importance of framing in experiments is well grounded. But accepting this insight doesn't imply that we must also believe that human beings are mindless robots, irreversibly programmed with rigid social behaviors. Given time and adequate incentives, we can learn by trial and error or by imitation to adapt our behavior to novel situations. Sometimes we even think a little about what we are doing.

Presumably the rate at which different people learn depends on their personal characteristics, and the strength of their conditioning in the social norm that they must learn to abandon. Perhaps some people will never learn, no matter how long we give them or how large the incentives. The study of such inflexible folk is certainly of very great interest. But the evidence from the one-shot Prisoners' Dilemma suggests that the inflexible fraction of the student population from which subjects are usually drawn can't be more than about 10 percent of the whole.

Fairness

Although game theorists like myself have to put up with being said to be unremmitingly hostile to the idea that fairness can influence human behavior, I have devoted a substantial chunk of my life to working out a theory of how and why fairness norms matter in human societies (Binmore 1994, 1998). I even have some lingering hope that the absence of any algebra in my recent *Natural Justice* will result in the theory getting some serious attention from moral philosophers (Binmore 2005).

The basic thesis of the theory is that our sense of fairness evolved because the coordination games of which everyday social life largely consists commonly have large numbers of equilibria. A society therefore needs equilibrium selection devices if its members are to succeed in coordinating on one particular equilibrium in each game. Fairness is our name for a class of equilibrium selection devices that result in some social surplus being divided.

The conclusions to which I am led accord rather well with a psychological literature referred to as "modern equity theory" that is largely ignored by economists.[4] This literature offers experimental support for Aristotle's ancient contention, in his *Nichomachean Ethics*, that what is fair is what is proportional.

I don't plan to press the virtues of my theory of fairness in this book, since I haven't done any experimental work of my own on the subject. But two aspects of this theory are immediately relevant here. The first is the significance of the theory of repeated games. The second is the importance of evolutionary theory.

Repeated Games

The folk theorem of repeated game theory says that any contract that rational players might sign on how to play a one-shot game is sustainable as an *equilibrium* outcome when the game is played repeatedly by patient players with no secrets from each other. Cooperative agreements that can only be sustained in one-shot situations with the assistance of an external enforcement agency can therefore survive as *self-policing* social norms in a repeated environment.

The mechanism that sustains self-policing cooperative agreements in repeated games is *reciprocity*. People sometimes register their understanding of how such self-policing agreements work by saying, "I'll scratch your back if you'll scratch mine." But such a promise wouldn't be effective without the implied threat that I'll stop scratching your back (or worse) if you stop scratching mine. That is to say, what keeps the cooperative arrangement on track is that everybody recognizes that they will suffer some punishment if they don't honor the implicit deal.

The need to punish deviant behavior is explicit when Adam and Eve both use the GRIM strategy in the infinitely repeated Prisoners' Dilemma. The GRIM strategy tells you to play *dove* at each repetition of the Prisoners' Dilemma until the opponent fails to reciprocate. After an opponent plays *hawk*, the GRIM strategy tells you to play *hawk* yourself ever after. Neither player can therefore profit from deviating from the GRIM strategy by being the first to play *hawk* because the deviant will be relentlessly punished by the opponent responding by always playing *hawk* thereafter.

When we all lived in small foraging communities, there was no external enforcement agency to police the way that people played coordination games, but most of the coordination games we played together were *repeated* day after day. Moreover, as in small villages today, everyone knew everyone else's business. Given the folk theorem of repeated game theory, it is therefore perhaps no great surprise that evolution—both cultural and biological—should have generated fairness norms that allow social surpluses to be divided efficiently in favorable environments without wasteful conflict (Axelrod 1984).

The conditions of the folk theorem don't apply in large modern states, but much of our interaction with other human beings nevertheless continues to be open-ended. Even when we won't be interacting with the same person again, the way we conduct ourselves with that person is often being observed by onlookers with whom we may well interact in the future. Punishment for cheating on a partner can then be administered not by the victim (as in the GRIM strategy) but by onlookers refusing to deal with someone who has just established a reputation for being untrustworthy. That is to say, the domain within which we may reasonably expect cooperation to survive as equilibrium behavior is much wider than the narrow class of games to which formal versions of the folk theorem apply directly.

For this reason I believe that the social norms to which we unconsciously appeal in bargaining and other social situations are often best thought of as being adapted to *repeated* interactions. Such cooperative norms for repeated games sometimes get triggered in one-shot laboratory situations. This would explain why inexperienced subjects commonly play *dove* in the one-shot Prisoners' Dilemma. But after getting shafted a few times when playing the one-shot Prisoners' Dilemma over and over again (against a new opponent each time) and finding themselves unable to retaliate, most people eventually shift to playing *hawk*.

Strong Reciprocity?

A recent anthropological study highlights how social norms can be triggered in the laboratory (Henrich et al. 2004, 2005). The study confirms that inexperienced citizens of different societies play a variety of canonical toy games in different ways—presumably reflecting the fact that different societies operate different social norms. As Henrich et al. (2005) say: "Experimental play often reflects patterns of interaction found in everyday life."

The anthropologist Jean Ensminger is more explicit when commenting on why the Orma contributed generously in the public goods game she carried out as part of the study:

> When this game was first described to my research assistants, they immediately identified it as the "*harambee*" game, a Swahili word for the institution of village-level contributions for public goods projects such as building a school. . . . I suggest that the Orma were more willing to trust their fellow villagers not to free ride in the Public Goods Game because they associated it with a learned and predictable institution. While the game had no punishment for free-riding associated with it, the analogous institution with which they are familiar does. A social norm

had been established over the years with strict enforcement that mandates what to do in an exactly analogous situation. It is possible that this institution "cued" a particular behavior in this game (Henrich et al. 2004, p. 376).

The enforcement here is operated by the players themselves as envisaged in the folk theorem, and not external enforcement operated by the government. (National or cross-regional attempts at *harambee* collections are predictably corrupt.)

Despite this and similar evidence from the anthropologists who contributed to the study, Henrich et al.'s (2004) introduction insists on interpreting the data as supporting the existence of significant other-regarding preferences. But if Ensminger is right, then it would be a huge mistake to try to explain the behavior of the Orma in her public goods game on the hypothesis that their behavior was adapted to the game they played in her makeshift laboratory. In particular, inventing other-regarding utility functions whose maximization would lead to generous contribution in the public goods game would be pointless. Ensminger is suggesting that the subjects' behavior is adapted to the public goods game embedded in the *repeated* game that they play every day of their lives, for which the folk theorem provides an explanation that does not require anything at all to be invented.

It is admittedly difficult to distinguish the interpretation of the data that I share with Ensminger from the claim that the subjects have the kind of other-regarding preferences postulated by the theory of "strong reciprocity." This theory holds that people have a liking for reciprocation built into their personal utility functions. I am always puzzled by the ardor with which advocates of the theory of strong reciprocity, like Bowles and Gintis (2002) and Gintis (2002), condemn the idea that people might also sometimes reciprocate favors because this is how cooperative equilibria are sustained in indefinitely repeated games. Don't they see that the folk theorem would provide a possible evolutionary explanation for the emergence of strong reciprocity? However, my guess is that they reject the support that the theory of repeated games might offer the strong reciprocity hypothesis because everyone can see that we don't need to hypothesize strong reciprocity if we can explain the available data without going beyond the so-called weak reciprocity used to prove the folk theorem.

Evolution?

Where did the fairness norms triggered in laboratory experiments come from? I believe they evolved as equilibrium selection devices for use in

those real-life games in which a surplus can be created by operating one of many cooperative equilibria. Cultural evolution must surely have been as important as biological evolution in this process, since what people regard as fair seems to depend heavily on both context and culture. Indeed I think that cultural evolution is active all the time in generating new social mini-norms for novel contexts. Some bargaining experiments can even be interpreted as snapshots of cultural evolution shaping a new fairness mini-norm while we watch (chapter 2).

But evolution is a slippery concept, easily harnessed in support of almost any doctrine. Other-regarding preferences are a case in point. It isn't good enough to argue that evolution built a regard for others into our preferences because we are all better off that way. The same argument shows that evolution should be expected to generate cooperation in the one-shot Prisoners' Dilemma. Similarly it isn't good enough to argue that evolution will select the preferences that we would choose to bind ourselves to if we knew our choices were to become common knowledge (Güth and Kliemt 1998). This is just another version of the Transparent Disposition Fallacy used by some authors in defense of rational cooperation in the one-shot Prisoners' Dilemma (Binmore 1994b). Any evolutionary defense for other-regarding preferences needs to be accompanied with a plausible story that explains *how* other-regarding mutants could have invaded our gene pool, and managed to survive once established—as, for example, in Samuelson (2004) or Weibull and Salomonsson (2005).

A Gift-Exchange Experiment

Nor can we afford to be naïve about evolutionary interpretations of laboratory experiments. An anecdote of Konrad Lorenz will serve to illustrate one particular mistake that I think it important to avoid.

Lorenz placed a totally inexperienced jackdaw on a marble-topped table, whereupon the baby bird went through all the motions of taking a bath. I think one may reasonably deduce that bath-taking behavior is genetically programmed in jackdaws, and that a trigger for this behavior is the presence of a flat, reflective surface (like water). What one isn't entitled to deduce is the absurd conclusion that bath-taking behavior somehow promotes the survival of jackdaws placed on marble-topped tables. If the jackdaw were human, we would say that its behavior was irrational, or ill-adapted to the context.

An example of the kind of interpretive mistake I am warning against is provided by a much-quoted experiment of Fehr et al. (1997) and Fehr and Gächter (2000). It can be thought of as modeling a competitive labor

market in which the workers have the opportunity to reward employers who pay above the competitive rate by putting in more effort—even though the employer has no comeback if the worker just pockets the extra money and shirks.

The finding is that workers do indeed reward generous employers with more effort—that they metaphorically "exchange gifts." The authors speculate that their data supports the theory of strong reciprocity, which says that people have preferences that incorporate a positive liking for reciprocity.

But before leaping to such a conclusion, shouldn't we consider a less dramatic scenario? Although the subjects are called buyers and sellers in the experiment rather than employers and workers, its framing nevertheless cues the subjects for the *repeated* environment typical of a labor market. It therefore triggers a fairness norm that selects one of the cooperative equilibria of such a repeated game. Reciprocity therefore matters to the behavior of the subjects because reciprocity is the mechanism that sustains cooperative equilibria in repeated games.

If this dull story is true, then instead of subjects responding *rationally* to a set of preferences unconsidered in traditional economics, they just have traditional preferences but are behaving *irrationally*, in the sense that their behavior isn't adapted to the one-shot game they are deemed to be playing in the laboratory.

Ledyard's (1995) survey of experiments on the Prisoners' Dilemma and related games is obviously relevant here. What would happen if the subjects in the Fehr et al. study were allowed to play a large number of times?

We have seen that it is uncontroversial that subjects in experiments change their behavior as they gain experience, and matters are no different in the current study. The observed movement is initially *away* from the behavior that the authors assume should be the orthodox equilibrium prediction. But who can say what would happen with more than the usual ten or so repetitions? Nevertheless, in summarizing their data, Fehr et al. (1997, p. 2) say (with my italics):

> These results indicate that reciprocity *motives* may indeed be capable of driving a competitive experimental market *permanently* away from the competitive outcome.

This claim is called into immediate question by the very data that the authors offer in its support. How could they have overlooked the final round effects evident in the data given in the appendix to their paper? In

16 of the 26 final rounds reported in which the worker has the opportunity to reciprocate, he doesn't. On the contrary, his effort is as small as it is possible for it to be.[5]

My own guess is that an understanding of what is really going on in the Fehr et al. experiment requires appealing to the contagion mechanism described by Kandori (1992) for sustaining cooperative equilibria in infinitely repeated games played by small groups of anonymous agents. It is true that the game of Fehr et al. is only repeated a finite number of times, but a number of authors, including Reinhard Selten (1986), have shown that the folk theorem often still works in the laboratory when the number of repetitions is finite. The fact that cooperation tends to break down in the final rounds of these experiment adds some support to my conjecture, once it is revealed that the same holds true in the experiment of Fehr et al. (chapter 8).

Social Preferences

When experimental economics was recognized in 2002 with a Nobel Prize awarded jointly to Daniel Kahneman and Vernon Smith, a joke circulated that Smith had been awarded the prize for showing that economics works in the laboratory, and Kahneman for showing that it doesn't.

The uncontroversial truth is that there are domains within which traditional economic theory—including game theory—works badly or not at all, and other domains within which it works rather well. What is controversial is how large these domains are, and where they lie.

Nowadays the followers of Daniel Kahneman and Amos Tversky[6] call themselves behavioral economists, to distinguish themselves from experimental economists like Vernon Smith or Charles Plott, who work largely in the tradition of neoclassical economics. However, on the subject of fairness in bargaining games there is a curious reversal of attitudes. Behavioral economists seem mostly to believe that the available experimental data support the hypothesis that laboratory subjects are classical optimizers whose utility functions have a social or other-regarding component.[7]

I have already explained why I think it a mistake to get into a dispute over what kind of utility function is being maximized by inexperienced and unmotivated laboratory subjects, but I want to insist that this doesn't imply that I believe that social preferences have no role to play in explaining human economic behavior in general. On the contrary, my own theory of fairness depends very heavily on the idea that social preferences

matter (Binmore 2005). The rest of this section is therefore an aside that briefly examines three different ways in which I believe that social preferences can be significant.

Blood Is Thicker Than Water

Hamilton's (1995) rule offers a biological prediction of the extent to which we should care about a relative. A gene that programs its animal host to maximize the gene's fitness would do best to take into account not only the children its current host might produce but also the children of the host's relatives. The probability that they will carry a copy of the gene is smaller but much too large to be neglected.

The point was famously made in a semi-serious joke of the biologist J. B. S. Haldane. When asked whether he would give his life for another, he replied that the sacrifice would only be worthwhile if it saved two brothers or eight cousins. Haldane's joke is only funny if you know that your degree of relationship to a full brother is one-half, and your degree of relationship to a full cousin is one-eighth. These numbers are the probabilities that a recently mutated gene in your body is also to be found in the body of the relative in question.

The only experimental study on Hamilton's rule of which I know found that best friends get pretty much the same consideration as brothers or sisters (Dunbar et al. 2004). My guess is that our bodies have to deduce their degree of relationship to others from the extent to which we find ourselves in their company. If so, then the instincts that promote altruism within the family may also be triggered within a sufficiently close-knit group of unrelated individuals, as in an army platoon under combat conditions or a teenage street gang.

This is perhaps why we find ourselves feeling curiously obligated to old school friends or office colleagues, whom we may actively dislike at the conscious level. Our bodies are telling us that this pushy individual demanding an inconvenient favor must be a cousin or an aunt—as she would probably have been when we all lived in small foraging communities. Even establishing eye contact with a beggar in the street somehow creates enough inner discomfort at neglecting a potential relative that we are sometimes moved to hand over our small change with no prospect of any recompense.

I therefore accept that most people have other-regarding preferences to some degree—that they are willing to pay a small amount for no other return than the warm glow they derive from improving the lot of another human being. Perhaps there are economists who think otherwise, but I

don't know who they are. One doesn't even need to appeal to the data from Dictator Games to confirm the claim, since nobody denies that nearly everyone contributes some small fraction of their income to charity. Moreover the kinship argument offers a possible evolutionary explanation of why people might be made this way. It is also doubtless true that some small fraction of people are willing to make large contributions on a regular basis toward the welfare of others, although an explanation of this behavior is not so easy to find.

However, the fact that some small fraction of the population behave like saints and that most of the rest of us are willing to treat pretty much anyone as a distant relative won't generate a warm enough glow to convert a game like the Prisoners' Dilemma into a game with an efficient equilibrium when the other player is a stranger. One needs *large* perturbations of the preferences economists traditionally attribute to players for this to happen. Matters are different in the games we play with the friends and neighbors in our extended family, but I don't believe the evidence offered in support of the claim that most of us are programmed to treat strangers like close members of the family survives serious examination.

Revealed Preference

Why do I reject the social preferences that behavioral economists fit to their experimental data? They commonly report relatively large warm-glow effects.

The theory of revealed preference tells us that we can describe the behavior of agents who choose consistently as optimization relative to some utility function. However, economists who take the orthodox neoclassical position seriously are very careful *not* to deduce that the observed behavior was generated by the agent *actually* maximizing whatever utility function best fits the data. This would be to attribute the kind of psychological foundations to neoclassical theory that its founders invented the theory to escape.

Being able to fit a utility function only tells us that the behavior is consistent—it doesn't tell us *why* the behavior is consistent. For example, one way of explaining the behavior of that half of the population of inexperienced subjects who cooperate in the one-shot Prisoners' Dilemma is to say that they are optimizing a social utility function whose arguments include the welfare of others. Another is to attribute any consistency in their behavior to the fact that they are unconsciously operating a social norm better adapted to repeated situations.

Both explanations fit the data equally well, but the former explanation is easier to criticize. What is the point of insisting that players have other-regarding utility functions built into their brains if doing so doesn't allow predictions to be made about how they will play in future, or in other games? But we know that the behavior of subjects in the one-shot Prisoners' Dilemma changes markedly over time as they pick up experience. A social utility function fitted to the behavior of an inexperienced subject will therefore fail to predict how he or she will behave when experienced—let alone when they play other games in other contexts.

None of this is to suggest that fitting utility functions to behavioral data may not be a useful way of summarizing the data—provided that we don't fall into the trap of assuming that the same utility function will necessarily predict other data without any experimental confirmation.

When evaluating an empirical claim that people have personal preferences with a large social component that has been quantified using experimental data, I therefore always ask myself what new data from other sources this claim has genuinely succeeded in predicting. I don't know of any cases at all that can be said to have unequivocally cleared this hurdle.

The theory of inequity aversion proposed by Fehr and Schmidt (1999) is usually quoted in denial of this skeptical assessment. (See chapter 4.) Fehr and Schmidt claim to have used data from ultimatum games to calibrate the parameters in the other-regarding utility function of their theory, and then used the calibrated utility function to predict the data from experiments on other games. However, Shaked (2005) has pointed out that this claim cannot possibly be true, because the data supposedly used to calibrate the parameters only restricts their range. When Fehr and Schmidt picked particular values of the parameters from within this range, they therefore made use of information that they should have denied themselves.[8]

Empathetic Preferences

Comparing utils across different individuals has been a controversial subject for a long time. Only recently have traditional economists stopped teaching the dogma that such interpersonal comparisons are intrinsically nonsensical. But how can fairness judgments be made if we have no way of comparing the welfare of those among whom a surplus is to be shared?

John Harsanyi (1977) invented a theory of interpersonal comparison of utility that makes good sense in the context of my theory of fairness (Binmore 2005). Harsanyi postulates social or empathetic preferences that

exist in parallel with the standard personal preferences with which we are all familiar. With some apparently mild assumptions, Harsanyi shows that such empathetic preferences can be summarized in terms of a rate at which Eve assesses Adam's personal utils relative to her own personal utils.

Empathetic preferences live in an entirely different world from personal preferences because their content is entirely hypothetical. For example, Eve expresses an empathetic preference when she says that she would rather be herself eating an apple than Adam wearing a fig leaf—but there is no way Eve is ever going to get the opportunity to swap bodies with Adam.

I think the reason that normal people are all capable of expressing such empathetic preferences is that we need them to assess who should get how much when using fairness norms as equilibrium selection devices. The internal process by which we make such judgments is largely a mystery to us, and so it isn't surprising that we often confuse our empathetic preferences with our more readily understood personal preferences—especially those personal preferences that capture our feelings about those close to us.

Psychologists avoid this confusion by separating the notion of empathy from that of sympathy. A confidence trickster may *empathize* with an old lady by putting himself in her position to see what tall tale is most likely to persuade her to part with her money. He may compare the distress that she will feel at the loss of her life savings with his own joy in having her money to spend. He may even need to brush a tear from his eye as he contemplates her plight. But he won't be diverted from swindling her unless he also *sympathizes* with her by including her welfare among the arguments of his personal utility function.

I think economists need to make the same distinction. I agree wholeheartedly with those behavioral economists who argue that fairness matters. I also agree that we can't make sense of fairness norms without some notion of a social preference. But we don't need to identify a social preference exclusively with a sympathetic preference. I believe that the social preferences to which we appeal when making fairness judgments are mostly empathetic preferences that implicitly describe the standard of interpersonal comparison to be applied.

Straw Men

Finally, I want to address the standard criticism that people like me have to face—that we fix our experiments to get results consistent with neoclassical economics.[9] This slander is often exacarbated by characteriza-

tions of neoclassical economics that belong in horror comics rather than serious academic studies.

For example, neoclassical economists are said to be wicked for supposedly putting around the theory that people are inherently selfish. There is even a small experimental literature in which students of economics are supposedly demonstrated to be more evil than other students (Frank, Gilovich, and Regan 1993). As a result I know of at least one case in which a university senate was asked to ban the teaching of rational choice theory on the ground that it is immoral!

I agree that politically motivated economists, both of the left and the right, often use phony arguments in support of immoral policies, but I am not politically active, and neither are most traditionally minded economists. We have no interest in defending the transparently wrong proposition that people are inherently selfish. Just like anyone else, we give money to charity and help old ladies cross the road. We don't run experiments to justify an irrational prejudice in favor of neoclassical economics. We run experiments to determine the domains within which the predictions of neoclassical economics work reasonably well.

When the predictions don't work in apparently favorable environments, we ask ourselves why. Sometimes the answer is that our test tubes need cleaning, and sometimes the answer is that the theory needs fixing. Much of the attention of young neoclassical theorists in recent years has correspondingly been devoted to trying to come up with theories of bounded rationality that explain laboratory behavior better than is possible for any optimizing theory, whether neo-classical or retro-classical. (See, for example, Rubinstein 1998.)

I do not understand why this modest research program attracts such ire from behavioral economists. Behavioral economics is now triumphant in its primary aim. Everybody agrees that we need to study microeconomic behavior empirically in both the field and the laboratory. Behavioralists therefore having nothing more to gain from dismissing those experimentalists who find that traditional economics sometimes works as dishonest apologists for a failed orthodoxy.

Karl Marx said that history repeats itself, first as tragedy and then as farce. But do we really need to repeat the history of suspicion and reproach that accompanied the controversy over cooperation in the one-shot Prisoners' Dilemma? Or the more recently defunct experimental controversy over expected utility theory?

It was the latter controversy that brought Kahneman and Tversky (1979) to prominence, along with behavioral economics. But where is

this controversy now? After much sound and fury, the exhausted combatants all seem to have retired from the field, leaving behind the consensus that all behavioral theories of how humans make decisions under risk are bad, but the least bad is traditional expected utility theory (Camerer and Harless 1994; Hey and Orme 1994).

Even if you are as sure about the failings of some other orthodoxy as Kahneman and Tversky were about expected utility theory, it may therefore still be worth your while to read papers that seem to defend the orthodoxy with a view to finding out what they actually say, rather than lending a credulous ear to those who attribute absurdly unrealistic beliefs to their unfortunate authors.