UNIVERSITÀ DEGLI STUDI DI FERRARA
DIPARTIMENTO DI ECONOMIA ISTITUZIONI TERRITORIO
Corso Ercole I d'Este, 44 – 44100 Ferrara

Quaderno n. 8/2006

April 2006

A Little Behavioralism Can Go a Long Way

Ken Binmore   Joe Swierzbinski

A Little Behavioralism
Can Go a Long Way

Ken Binmore
Economics Department
University College London
Gower Street
London WC1E 6BT UK

Joe Swierzbinski
Business School
University of Aberdeen
Dunbar Street
Old Aberdeen AB24 3QY, UK

Abstract

This paper considers the data from several experiments, including both
the Ultimatum Game and a new experiment on Rubinstein's bargaining
model. It is shown that one does not need to follow the behavioral econo-
mists in attributing large deviations from predictions based on neoclassical
reasoning with money payoffs in order to explain the data.

A Little Behavioralism
Can Go a Long Way

by Ken Binmore and Joe Swierzbinski

1. Economic Man Refuted?

There is a school of behavioral economists who have popularized the notion that the neoclassical paradigm of *homo economicus* is refuted by the experimental evidence. We agree that the idea that human behavior can always be modeled as the rational optimization of money rewards in each and every context is off the wall, but who would want to defend such a wild claim? To make their case, behavioral economists need to address the more moderate claim that people often learn to play like income maximizers—given sufficient time and adequate incentives.

It isn't enough to look only at the behavior of inexperienced subjects. Nobody denies that they are unlikely candidates for the role of economic man. Nor is it enough to keep pointing at unusual games like the Ultimatum Game, in which subjects do not seem to adjust their behavior much as they gain experience. Indeed, it seems palpably dishonest to harp continually on such games, while simultaneously turning a blind eye to the very much larger literature in which laboratory subjects are reported as converging on the Nash equilibria of games with money payoffs.

Why do we see apparently anomalous behavior in the class of games to which behavioral economists restrict their attention? This paper argues that the explanation lies partly in the fact that behavioral economists are some twenty years behind the times in thinking that economic man must solve games using the principle of backward induction, whereas advances in evolutionary game theory have shown that it is unwise to discard any Nash equilibrium whatever without close attention to the context. The paper then goes on to explore the extent to which the set of Nash equilibria in some games that behavioral economists regard as canonical can be expanded by deviating only slightly from the income-maximizing hypothesis.

The same approach is then applied to an experiment of our own on the Rubinstein bargaining game [20] with unequal discount rates. A full discussion of the experimental details and an analysis of the results is given elsewhere (Binmore *et al* [11]). The results are supportive of the pure Rubinstein prediction in some contexts but not in others.

The laboratory successes enjoyed by the Rubinstein theory this experiment and others (Binmore *et al* [9, 8], Camerer [12]) may seem paradoxical to those who believe that the theory stands or falls on whether human subjects commonly use backward induction in the laboratory. However, it should be noted that

the unique subgame-perfect equilibrium of Rubinstein's model also happens to be a Nash equilibrium with stationary expectations. One can therefore find an evolutionary explanation for why subjects sometimes find their way to the Rubinstein prediction that does not require defending the discredited principle of backward induction.

We pursue this evolutionary explanation here by applying it to the contexts in which the pure Rubinstein theory is not very successful in predicting the behavior of laboratory subjects. For these contexts, the Rubinstein bargaining model with unequal discount factors needs to be added to the class of anomalous cases identified by behavioral economists. Our purpose is to observe that the anomalies can largely be accommodated by assuming that some fraction of the population of subjects are slow in learning that the fair outcome on which they may have been conditioned in the past isn't adapted to the game they are playing in the laboratory.

In summary, we believe that behavioral economists are right to argue that the income-maximizing hypothesis for experienced and adequately incentified subjects needs to be modified to accomodate anomalous cases, but that it is unproven that there is a need for the modifications to be large in the kind of bargaining situations we have studied. The issues are pursued at greater length in a forthcoming book (Binmore [5]), on chapter 8 of which the current paper is based..

2. Ultimatum Game

When subjects first encounter a new game in the laboratory, we do not believe that they commonly recapitulate the principles of game theory in their heads and play accordingly. We therefore do not believe that the subjects are actively optimizing relative to any utility function whatever, whether other-regarding or selfish. We think instead that inexperienced subjects respond to the framing of the experiment by playing according to whatever social norm is triggered by the hints and cues with which they are presented. Game theory is relevant to such social norms, because we believe they evolved in the first place as equilibrium selection devices for the *repeated* games of everyday life.

But human beings are not helpless robots, irrevocably programmed by their culture with fixed behaviors. We vary in our flexibility when confronted with new situations, but most of us can and will learn if given the opportunity, and the vast majority of relevant experiments confirm that subjects move towards a Nash equilibrium—calculated with money payoffs—of the laboratory game they are playing, provided that adequate incentives and sufficient time for learning are built into the experimental design.

However, there is a minority of anomalous cases in which subjects do not shift much or at all from their initial behavior. How is such behavior to be explained? Behavioral economists offer the explanation that they are already at or close to a Nash equilibrium of a game in which their payoffs are not measured in money,

but in units of utility that take into account of the welfare of other players or other social considerations. We agree that one can explain the anomalous cases by arguing that the players are already at or close to a Nash equilibrium, but we see no need to modify the assumption that subjects maximize expected money by very much in order to make this explanation work.

The reason that one doesn't have to move far (or sometimes at all) from the income-maximizing hypothesis to explain the anomalous cases is that the games involved typically have large numbers of Nash equilibria that behavioral economists neglect to take into account. If the social norm that is triggered by the way an experiment is framed happens to coordinate the behavior of the subjects on or near one of these neglected equilibria, then any learning that follows will have little effect. The subjects will not be led away to a distant Nash equilibrium, because they are already in the basin of attraction of a nearby Nash equilibrium.

The leading anomalous case is the Ultimatum Game. In the Ultimatum Game, a sum of money can be divided between Alice and Bob if they can agree on a division. The rules are that Alice proposes a division and that Bob is then restricted to accepting or refusing. If the subgame-perfect equilibrium (in which Bob acquiesces when Alice demands almost all the money) were the only Nash equilibrium of the game, then the fact that Alice's modal offer in the laboratory is a fifty:fifty split would be a serious challenge to the income-maximizing hypothesis for experienced players, since this conclusion seems to be robust when the amount of money is made large or repeated play (against a new opponent each time) is allowed.

However, as with other anomalous cases, the Ultimatum Game actually has many Nash equilibria. In fact, any split of the money whatsoever is a Nash equilibrium outcome on the income-maximizing hypothesis. Not only does the Ultimatum Game have many Nash equilibria, but computer simulations show that simple models of adaptive learning can easily converge on one of the infinite number of Nash equilibria that are not subgame-perfect (Binmore, Gale and Samuelson [6]).

However, this isn't the point of presenting the computer simulation illustrated in Figure 1, which was one of a large number of simulations carried out for Binmore, Gale and Samuelson [6]. In this simulation, the original sum of money is $40 and the simulation begins with Alice offering Bob about $33, leaving $7 for herself. One has to imagine that the operant social norm in the society from which Alice and Bob are drawn selects this Nash equilibrium outcome from all those available when ultimatum situations arise in their repeated game of life. This split (like any other split) is also a Nash equilibrium outcome in the one-shot Ultimatum Game.

The figure shows our slightly perturbed replicator dynamic leading the system away from the vicinity of this $(7, 33)$ equilibrium. The system eventually ends up at a $(30, 10)$ equilibrium. The final equilibrium isn't subgame-perfect (where the split would be $(40, 0)$), but this fact isn't particularly germane. What
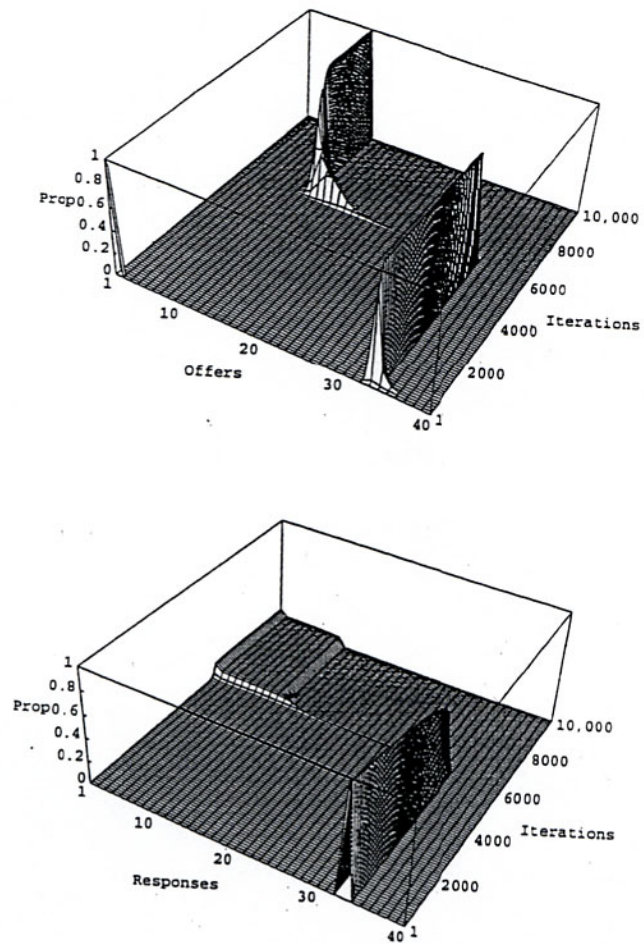
4

Figure 1: Simulated adaptive learning in the Ultimatum Game. The upper diagram shows the evolution over time of the offers a large population of player I's would make to player II if chosen to play. The diagram on the right shows the evolution over time of the acceptance levels of a corresponding population of player II's. A slightly perturbed version of the replicator dynamics is simulated whose parameters have been chosen to make the system converge on a 30:10 split of the $40 available. This takes 5,000 or so iterations when the system is started close to a 7:33 split. (The suddeness of the eventual transition between the Nash equilibria at 7:33 and 30:10 is illusory as the number of iterations during the transition exceeds by far those in any Ultimatum Game experiment.)

is important here is that it takes some 5,000 periods before our simulated adaptive process moves the system any significant distance from the vicinity of the original $(7, 33)$ equilibrium. This enormous number of periods has to be compared with the 10 or so commonly considered "ample" for adaptive learning to take place in the laboratory.

One might summarize these remarks by saying that testing the Ultimatum Game isn't an ideal way to go about exploring the extent to which an income-maximizing version of game theory works. Any efficient deal corresponds to a Nash equilibrium that a social norm operating in the society from which the subjects are drawn might render focal. A suitably perturbed adjustment process might eventually lead the subjects elsewhere, but the number of iterations this is likely to take would not be easy to replicate in a laboratory.

3. Public Goods Games with Punishment

In games like the Prisoners' Dilemma that can be interpreted as modeling the private provision of public goods, it is uncontroversial that most experienced subjects in laboratory experiments contribute little or nothing. However, Fehr and Gächter's [13] show that the situation changes when free riders can be punished after the contribution phase is over.

In their modified experiment, the subjects can pay a small amount to reduce the payoff of a free rider of their choice by a substantially larger amount. The opportunity to punish free riders in this way is actually used by the subjects, although an income-maximizer can gain nothing from such behavior. Contributions correspondingly rise progressively until most subjects are contributing a substantial amount. The conclusion drawn is that the subjects have a liking for punishing defectors built into their utility functions.

It is doubtless true that most people are disposed to punish anti-social behavior even when there is no money to be made out of this practice. But how firm is this tendency? Will more experience teach people that they gain nothing from punishing malefactors whom they will never meet again? How much of a loss will people endure before giving up the opportunity to punish?

Fehr and Gächter's [13] experiment is uninformative, because they overlook the fact that attributing only the trace of a liking for punishing bad behavior to the subjects is enough to create a Nash equilibrium in their game in which everybody contributes maximally (Steiner [22]). Each player's strategy in this equilibrium calls only for the worst free rider to be punished. Since all players punish the worst free rider, their share of the cost of providing an adequate disincentive becomes tiny. However, the assumption that players are prepared to pay this tiny cost is adequate to support the equilibrium, because nobody wants to be the worst free rider.

4. A Gift Exchange Game

6

An experiment of Fehr *et al* [15, 14] is based on an idealized competitive labor market in which the workers have the opportunity to reward employers who pay above the competitive rate by putting in more effort. Subjects representing workers turn out to reward generous employers with more effort, although the employers have no way of identifying workers who shirk with a view to punishing them in the future. The result is typical of "gift-exchange" experiments that are offered in support of the hypothesis that people have preferences that incorporate a positive liking for reciprocating.

In a simplified version of the kind of labor market studied in this literature, there are m employers and n workers, where $m < n$. Each of N periods begins with each employer independently publishing either a *high* wage or a *low* wage for all to see. The workers get a negative payoff from being unemployed, and so they compete to get employed. Each worker has an equal chance, and so the probability that any single worker finds employment in any given period is m/n. The matchings are entirely anonymous, so that long-term relationships between an employer and a worker are impossible.

A worker on a *low* wage automatically shirks. But a worker on a *high* wage can choose *high* or *low* effort. Both members of a matched pair receive a payoff of s if the wage is *low* (and so the worker shirks). Both receive b if the wage is *high* and the worker puts in *high* effort. The worker receives a payoff of 1 and the employer a payoff of 0 if the wage is *high* and the worker puts in *low* effort. We assume that $0 < s < b < 1$.

All Nash equilibria of this finitely repeated game require that the employer offers a *low* wage along the equilibrium path, but matters change if the game is perturbed slightly. To this end, we assume that each player is independently strategic with probability $1 - \pi$, or a reciprocating robot with probability π. A reciprocating robot makes a *high* offer as an employer and puts in *high* effort when receiving a *high* wage as a worker—until he observes that anyone at all has deviated from this behavior, after which he always plays *low*. The strategic players do not know the value of π, but update their subjective probability distribution for this parameter as play proceeds.

For small values of π and large enough values of N, there are Nash equilibria of this finitely repeated game in which everybody plays *high* until near the end of the game. Cooperation is sustained by the contagion mechanism identified by Kandori [16] for infinitely repeated games. The game is only finitely repeated, but the introduction of a small fraction of reciprocating robots permits a similar cooperating equilibrium to be sustained. As in the gang-of-four paper of Kreps *et al* [17], strategic players find it expedient to mimic the robots until it no longer matters whether a robot is provoked into precipitating a breakdown.

A number of authors, including Reinhard Selten, [21] 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 experiments adds some support to the relevance of the preceding model, since the same holds true in the experiment of Fehr *et al* [15], with 16 out of 26

workers putting in only the minimum effort in the tenth and final round.

5. Bargaining with Unequal Discount Rates

An experiment on Rubinstein's [20] bargaining model with unequal discount rates reported elsewhere supports the hypothesis that most subjects optimize to a degree that would eventually be sufficient to shift a group of experimental subjects to the Rubinstein solution if all members of the group were to behave in the same way (Binmore, Swierzbinski and Tomlinson [11]). But some subjects presumably do not learn so quickly as others. Perhaps some do not learn at all, but remain fixated on operating what they regard as a fair social norm. If we perturb Rubinstein's model by writing such behavioral possibilities into his scenario, what impact will this have on the predicted outcome?

In seeking an answer to this question, we focus on models in which a fraction of the population of possible players are initially conditioned on an outcome f of the bargaining problem, which they regard as fair or focal. However, their behavior is not inflexible. After observing a refused proposal, they sometimes switch to playing strategically with some exogenously determined probability. We find that the existence of such a group can result in significant perturbations of the Rubinstein outcome—even when all the conditioned players will eventually end up playing in the same way as the strategic players.

5.1. Experimental Background

This section briefly reviews some relevant experimental evidence.

Subgame perfection?   The experimental evidence on finite bargaining games with alternating offers is firmly hostile to the idea that laboratory subjects use backward induction in deciding how to play (Camerer [12]). Even when it is assumed that the players care about their opponent's payoffs as well as their own, backward induction performs badly (Binmore *et al* [7]).

It is therefore commonly thought that Rubinstein's use of the concept of a subgame-perfect equilibrium in analyzing his infinite-horizon model makes his theorem irrelevant to the behavior of real people. However, in the case of equally patient players, it turns out that the Rubinstein theory does rather well in predicting experimental outcomes when compared with more traditional bargaining approaches (Binmore *et al* [9, 8]). One possible explanation is that the conclusion of Rubinstein's theorem doesn't change if we replace the idea of a subgame-perfect equilibrium by that of a stationary expectations equilibrium— to which subjects may perhaps find their way in repeated play using some kind of myopic adjustment procedure, in which tomorrow is always treated as though it will resemble today.

Learning?   Although the experimental evidence that laboratory subjects can adjust their behavior over time to the strategic realities of most simple games is overwhelming, the case of finite bargaining games with alternating offers is more problematic, with nearly all experiments finding little or no evidence of experience changing the subjects behavior (Camerer [12]).

However, in the bargaining games we have studied experimentally, we have always found evidence of learning—sometimes very rapid learning—provided that the feedback provided is sufficiently rich. A possible explanation is that simple models of trial-and-error adjustment in the Ultimatum Game (and so presumably in similar games) predict that any learning is likely to be painfully slow (Binmore *et al* [6], Roth and Erev [19]).

Atypical subjects.   A particularly strong body of evidence is presented by Ledyard [18] in his survey of a very large number of games like the Prisoners' Dilemma that model the private provision of public goods. Novices cooperate somewhat more than half the time, but the frequency of cooperation declines as the subjects gain experience, until about 90% of the subjects are defecting. However, the remaining 10% of the subject pool is of very considerable interest, especially since we find a similar proportion of subjects in our own bargaining experiments who seem impervious to strategic considerations (Binmore *et al* [8]).

A recent experiment.   In our most recent experiment, subjects played a variant of Rubinstein's [20] bargaining game in which the next proposer after a disagreement is chosen at random (Binmore *et al* [11]). The disagreement point is located at the origin. The feasible set resembles that of Figure 2. The subjects played a total of 24 games, sometimes as player I and sometimes as player II.

The subjects first knowingly played 8 "practice" rounds against a computer programmed to try to condition them either on the approximately utilitarian outcome $(8, 2)$, or on the equal-increments or Rawlsian outcome $(4, 4)$. They then knowingly played 16 times against other subjects in their group, chosen unpredictably anew at the start of each game. Our intention was to study the extent to which the stability of any focal points established by the conditioning is related to the location of the Rubinstein solution (Binmore *et al* [10]).

When player I's discount factor is $\delta_1 = 0.9$ and player II's is $\delta_2 = 0.8$ in the bargaining problem of Figure 2, the utilitarian outcome $(8, 2)$ is the Rubinstein solution. When the players' discount factors are exchanged, the Rawlsian outcome $(4, 4)$ becomes the Rubinstein solution. Introducing one or other of these pairs of discount factors allows four treatments to be distinguished:

Treatment 1 Subjects conditioned on $(4, 4)$. Rubinstein solution $(4, 4)$.

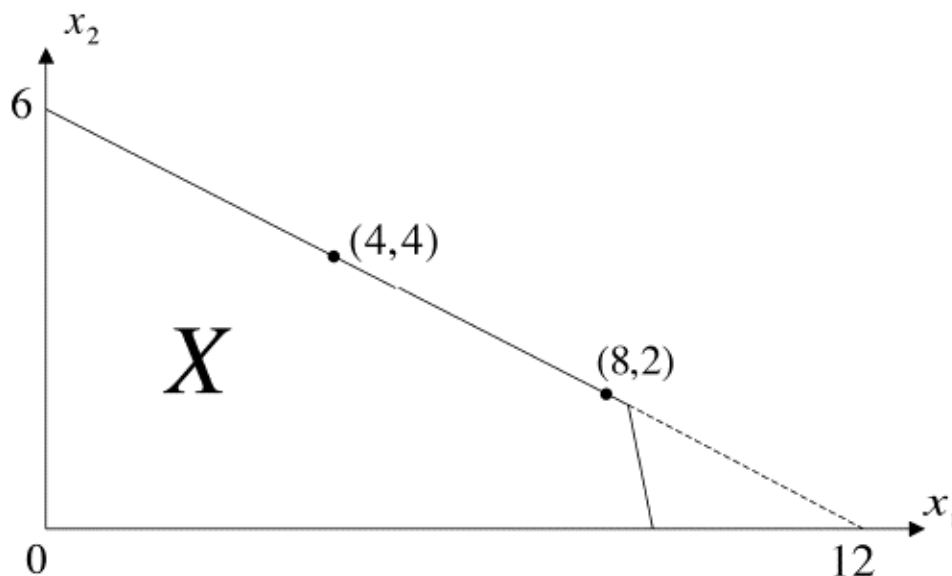**Treatment 2** Subjects conditioned on $(8, 2)$. Rubinstein solution $(4, 4)$.

Figure 2: A simplified version of the feasible set used in Binmore *et al's* [11] experiment on Rubinstein's Alternating Offers Game.

Treatment 3 Subjects conditioned on $(4, 4)$. Rubinstein solution $(8, 2)$.

Treatment 4 Subjects conditioned on $(8, 2)$. Rubinstein solution $(8, 2)$.

We succeeded in conditioning the subjects on $(4, 4)$, but we only succeeded in persuading the subjects that player I should get something more than 7 when our target was $(8, 2)$. (The computer randomized over a small range centered on $(8, 2)$. The subjects responded by moving close to an optimal response to this behavior.)

Some of our results are shown in Figures 3 and 4. The horizontal axis shows the 16 games played against a human opponent. The plus signs indicate that player I made the first proposal in a game. The vertical axis shows the mean monetary payoff to player I. The points marked with a cross show the mean payoff to player I in the final agreement, discounted to time zero (in each game). The stars indicate the mean amount assigned to player I by the opening proposal. The circles show the predictions of a myopic best-reply model. (Responders were shown the outcome of six recent final agreements discounted to the next period that they could use in estimating a best reply.) The squares show the predictions of a version of the myopic best-reply model which has been perturbed by introducing a small bias towards the equal-split outcome $(4, 4)$.

The fact that the mean initial proposals always recognize the strategic advantage of the proposer suggests that the Rubinstein approach is basically on track, but the steady movement toward the Rubinstein solution in Treatment

2 is absent (or only very slight) in Treatments 3 and 4. Why does player I not make more aggressive proposals in Treatment 3, and so further shift the trajectory of final agreements toward the Rubinstein solution of $(8, 2)$? Why again does the trajectory in Treatment 4 not shift from around $(7, 2\frac{1}{2})$ toward the Rubinstein solution of $(8, 2)$?

We do not see how it is possible to answer such questions in our experiment simply by attributing social preferences to the subjects that lead them to play fair. The data shows that the subjects' behavior varies so much over time that any such preferences would sometimes need to be malleable to an extent that would render worthless any attempt to describe the subjects' behavior exclusively in such terms. Their conditioning, their role in the game, and their experience of previous play evidently all matter a great deal. In particular, Treatment 2 shows clear evidence of learning—not only within each game—but between games as well.

On the other hand, fairness considerations are clearly relevant to our data—as they are in all bargaining experiments of which we are aware. However, there is an alternative explanation for why people sometimes play fair to the claim that a strong propensity for such behavior is frozen into their preferences. It is that fairness norms evolved as equilibrium selection devices (Binmore [3, 4]). It is this alternative explanation that motivates the model explored in this paper.

6. Perturbing Rubinstein's Model

In Rubinstein's [20] Alternating Offers Game, two players alternate in proposing how to split a shrinking cake. We model the cake at time 0 as the set

$$X_0 = \{x \in \mathbb{R}^2 : x_2 \leq g(x_1)\},$$

where $g : \mathbb{R} \to \mathbb{R}$ is strictly decreasing and concave. Its inverse function is denoted by $h : \mathbb{R} \to \mathbb{R}$. The set of Figure 2 will be used as a canonical example. This is really the special case when the boundary of X_0 is $x_1 + 2x_2 = 12$, since the chunk cut away from this set in Figure 2 is irrelevant to any calculations—although not to focal point considerations, as the midpoint $(6, 3)$ on the hypotenuse would clearly have strong focal properties if the cutaway chunk were present.

In order that our subjects need only look one move ahead in computing a stationary expectations equilibrium, our experiment modified the rules of Rubinstein's game so that the new proposer is always decided by the fall of a fair coin. This change does not alter Rubinstein's conclusions in any essential way.

At each time $t = 0, 1, 2, \ldots$ that the modified game is still in progress, an independent chance move chooses player I or II with equal probability to act as proposer or responder at this time. The proposer then makes a demand that the responder can accept or refuse. If the demand is accepted, the proposer receives his demand, and the responder is assigned whatever remains of the cake.
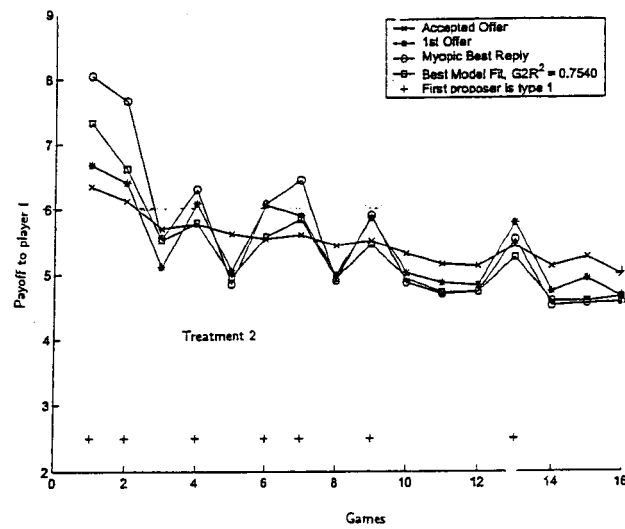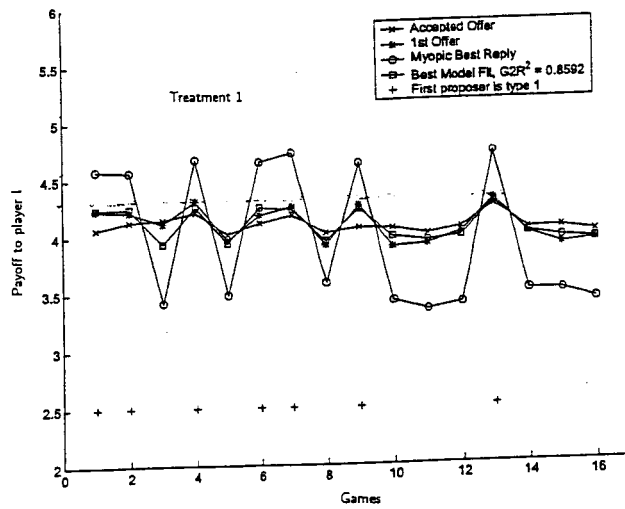
Figure 3: Data from Treatments 1 and 2 in Binmore *et al's* [11] experiment on a Rubinstein bargaining game.
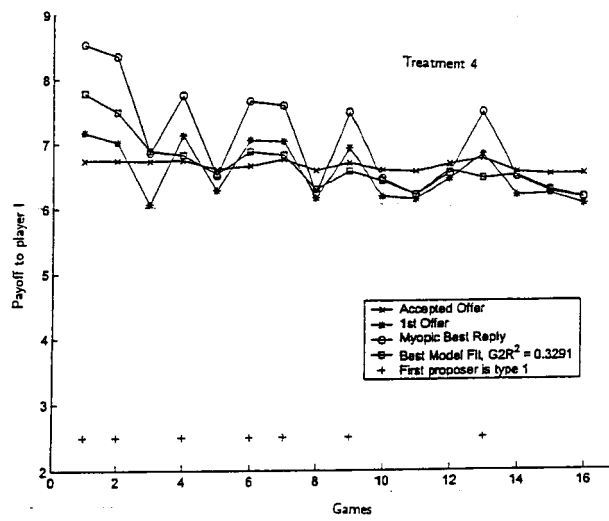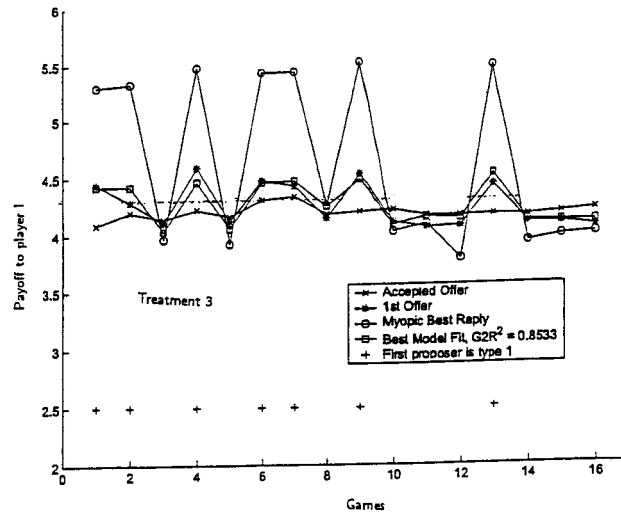
Figure 4: Data from Treatments 3 and 4 in Binmore *et al's* [11] experiment on a Rubinstein bargaining game.

The shrinkage of the cake is modeled by assigning discount factors δ_1 and δ_2 to the two players. After a refusal at time t, the cake shrinks from X_t to

$$X_{t+1} = \{x \in \mathbb{R}^2 : (x_1\delta_1, x_2\delta_2) \in X_t\}.$$

Since we assume $0 < \delta_i < 1$, the cake shrinks to zero if all proposals are refused.

The game has a unique subgame-perfect equilibrium (Binmore [2]). In equilibrium, the expected payoffs to the two players in our canonical example are

$$r_1 = \frac{12(1 - \delta_2)}{2 - \delta_1 - \delta_2}; \qquad r_2 = \frac{6(1 - \delta_1)}{2 - \delta_1 - \delta_2}. \tag{1}$$

Such computations are eased by noting that the answer turns out to be a stationary expectations equilibrium. Since a proposer will always make an offer (either $\delta_1 r_1$ or $\delta_2 r_2$) that leaves the responder indifferent between accepting and refusing, we merely need to solve the equations:

$$\begin{aligned}
2r_1 &= h(\delta_2 r_2) + \delta_1 r_1, \\
2r_2 &= g(\delta_1 r_1) + \delta_2 r_2.
\end{aligned}$$

Robots. Abreu and Gul [1] have studied the Rubinstein bargaining model in the case when it is common knowledge that there is some probability that an opponent will turn out to be a robot who always plays "fair" regardless of the strategic situation. As in the gang-of-four model (Kreps *et al* [17]), they find that a rational player will then sometimes pretend to be such a robot until some randomly determined number of proposals have been refused.

There are two reasons why we do not appeal to the Abreu-Gul model in seeking to make sense of our experimental data. The first is that it seems unlikely that their equilibrium could easily be learned by real people under laboratory conditions. The second is that our experience suggests that even strategically unresponsive subjects are a lot less inflexible than the robots of their model. Our own simpler model seeks to make virtues out of these problems.

Instead of a single chance move that decides whether a player will be a robot or a strategist at the start of the game, we introduce independent chance moves immediately following each refusal that permanently transform a player who has been a robot hitherto into a strategist from now on with probability $1 - \theta < 1$. In this way, we hope to capture in a crude way the fact that subjects who have been conditioned to play fair have the capability of learning to behave otherwise. If we keep things simple by always assigning the same belief to a newly created strategist as any other strategist would have on reaching the same point in the game, we simultaneously create a game with a stationary structure. Stationary expectations equilibria of this game then have a chance of being learned by subjects who operate some kind of myopic optimization process.

This specification leaves open the initial probability $\phi > 0$ that a player is a robot. In the calculations that follow, we take $\phi = \theta$ to keep things simple (but

see Section 7). It also leaves open the definition of a robot, which we take to be a player who has been conditioned to believe that the correct proposal is some efficient point f of X_0. A robot in the role of player I therefore always demands f_1 when proposing, and accepts f_2 or better when responding. A robot in the role of player II always demands f_2 when proposing, and accepts f_1 or better when responding.

Types of equilibrium. The plan is to investigate equilibria in which strategists always accept proposals made in equilibrium by strategists. Any refusal therefore signals to a strategist that the opponent is currently a robot, who will remain a robot only with probability θ in the next round. We can therefore employ the same methodology used to characterize stationary expectations equilibria in the unperturbed model. The only difference is that now a proposer sometimes has two possibly optimal demands to compare: a larger demand that makes a strategic responder indifferent between accepting and refusing, and a possibly smaller demand that will also be accepted by a robot responder.

We distinguish three types of equilibrium:

Rubinstein equilibria: A strategist always makes a demand that renders another strategist indifferent between accepting and refusing. In equilibrium, strategists always accept.

Fair equilibria: A strategist always makes the fair demand. In equilibrium, strategists always accept.

Hybrid equilibria: A strategist plays as in a Rubinstein equilibrium or as in a fair equilibrium, depending on whether assigned the role of player I or player II. In equilibrium, strategists always accept.

In designing our experiment, we did not contemplate equilibria other than those of the Rubinstein type, nor did we realize that the existence of a robot fringe could significantly alter the players' behavior in such equilibria. We now think that only the results in Treatment 2 look like the subjects are moving toward an equilibrium of the Rubinstein type. In the case of Treatment 1, we should have been ready to see a fair equilibrium with $f = (4, 4)$. In Treatments 3 and 4, we should have been ready to consider hybrid equilibria.

The point here is not to argue that one or other of these equilibria should be used to predict the data. We think that the modified myopic best-reply model used in Binmore *et al* [11] is to be preferred for this purpose, because it takes better account of the fact that even strategically minded folk need to learn to play equilibria. The point is rather that critics who would like to argue that the Rubinstein theory is altogether refuted by the data need to look harder at possible variants of the theory before they settle on such a draconian conclusion.

Rubinstein equilibria. Let r be the payoff pair strategic players expect before the game begins. We distinguish three cases.

$$\text{Case 1}: \ f_1 \ > \ \delta_1 r_1\,; \ f_2 > \delta_2 r_2\,.$$
$$\text{Case 2}: \ f_1 \ > \ \delta_1 r_1\,; \ f_2 < \delta_2 r_2\,.$$
$$\text{Case 3}: \ f_1 \ < \ \delta_1 r_1\,; \ f_2 > \delta_2 r_2\,.$$

In Case 1, a robot always refuses a strategist's offer of $\delta_2 r_2$ or $\delta_1 r_1$. When a strategic player I proposes, he therefore expects $(1-\theta)h(\delta_2 r_2) + \theta \delta_1 r_1$. Strategists always accept offers made by strategists, and so expect $\theta f_1 + (1-\theta)\delta_1 r_1$ when responding as player I. Similar considerations apply to strategic player IIs. The characterizing equations for a Rubinstein equilibrium in Case 1 are therefore:

$$2r_1 \ = \ (1-\theta)h(\delta_2 r_2) + \theta f_1 + \delta_1 r_1\,,$$
$$2r_2 \ = \ (1-\theta)g(\delta_1 r_1) + \theta f_2 + \delta_2 r_2\,,$$

These equations apply if and only if:

$$(1-\theta)h(\delta_2 r_2) + \theta \delta_1 r_1 \ \geq \ f_1\,,$$
$$(1-\theta)g(\delta_1 r_1) + \theta \delta_2 r_2 \ \geq \ f_2\,,$$

since it would not otherwise be optimal for strategists to tolerate their offers being refused by robots.

In Case 2, a robot in the role of player II accepts a a strategist's offer of $\delta_2 r_2$. When a strategic player I proposes, he therefore expects $h(\delta_2 r_2)$. A strategic player II refuses a fair offer, and so expects $\delta_2 r_2$ when responding. The characterizing equations for a Rubinstein equilibrium in Case 2 are therefore:

$$2r_1 \ = \ (h(\delta_2 r_2) + \theta f_1 + (1-\theta)\delta_1 r_1\,,$$
$$2r_2 \ = \ (1-\theta)g(\delta_1 r_1) + +(1+\theta)\delta_2 r_2\,.$$

These equations apply if and only if:

$$h(\delta_2 r_2) \ \geq \ \theta f_1 + (1-\theta)\delta_1 r_1\,.$$
$$(1-\theta)g(\delta_1 r_1) + \theta \delta_2 r_2 \ \geq \ f_2\,,$$

Case 3 is the same as Case 2, except that the roles of players I and II are reversed.

Fair equilibria. Fair equilibria can only exist in Case 1, because then $r = f$. The inequalities that need to be satisfied are:

$$f_1 \ \geq \ (1-\theta)h(\delta_2 r_2) + \theta \delta_1 r_1\,,$$
$$f_2 \ \geq \ (1-\theta)g(\delta_1 r_1) + \theta \delta_2 r_2\,.$$

These inequalities always hold when f coincides with the Rubinstein outcome in the unperturbed game and $\theta \geq \frac{1}{2}$. (There are no other fair equilibria in our canonical case when $\theta = \frac{1}{2}$.)

In particular, if $\theta \geq \frac{1}{2}$ and $f = (4, 4)$, it is an equilibrium in our Treatments 1 and 2 ($\delta_1 = 0.8$ and $\delta_2 = 0.9$) for everyone always to propose and accept the outcome f. The same holds in our Treatments 3 and 4 ($\delta_1 = 0.9$ and $\delta_2 = 0.8$) with $\theta \geq \frac{1}{2}$ and $f = (8, 2)$.

Hybrid equilibria. We omit the characterization of hybrid equilibria, since it will now be evident how this proceeds.

Existence. In our canonical example, computerized calculations reveal that one of these three types of equilibria exists for all values of θ ($0 \leq \theta \leq 1$) and all values of f ($0 \leq f_1 \leq 12$). There are occasionally multiple equilibria, but mostly only one of the three types of equilibrium exists for each pair (θ, f).

When the two parameters θ and ϕ are not equal, it becomes more complicated to characterize the equilibria. However, computerized calculations again show that one of the three types of equilibrium always exists, except for a few patches in the parameter space. The equilibrium is again typically unique.

7. What do Perturbed Equilibria Look Like?

Figures 5 and 6 show equilibrium behavior in perturbed versions of Rubinstein's model. They are directly comparable with the experimental data illustrated in Figures 3 and 4. In particular, the choice of who makes the first proposal in each game is exactly the same.

The firm lines in Figures 5 and 6 join points that show the average money payoff to player I in the final agreement, discounted to time zero (in each game). The broken lines join points which show the average money payoff proposed for player I at the outset of each game.

Notice that Treatment 1 in Figure 5 is a fair equilibrium in which both the firm and the broken graph sit on top of each other. Treatment 2 in Figure 6 is a Rubinstein equilibrium. Treatment 3 in Figure 5 is a hybrid equilibrium. Treatment 4 in Figure 6 is begins as a hybrid equilibrium, but switches to a Rubinstein equilibrium when the remaining fraction of robots becomes sufficiently small.

The parameter value $\phi = 0.5$ (which gives the fraction of robots at the beginning of the game) was chosen to correspond roughly with the fraction of subjects who begin by cooperating in Prisoners' Dilemma experiments. The parameter value $\psi = 0.1$ (which was taken to be zero in the previous section) is the fraction of robots who are assumed never to alter their conditioned behavior under any circumstances. This was chosen to correspond roughly with the fraction of subjects who persist in cooperating in Prisoners' Dilemma experiments after having
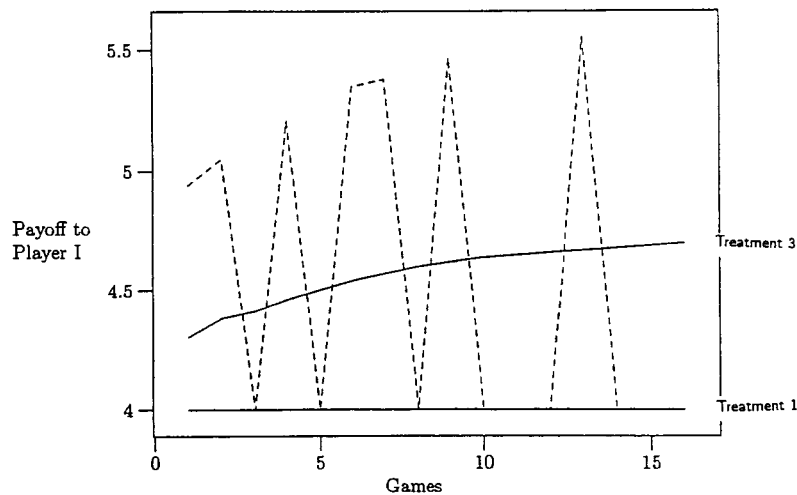
Figure 5: Equilibrium behavior in a perturbed Rubinstein model for Treatments 1 and 3. The unperturbed Rubinstein outcomes are $(4,4)$ and $(8,2)$ respectively. The parameters of the model are $\theta = 0.6$, $\phi = 0.5$, $\psi = 0.1$, and $f = (4,4)$. Treatment 1 is a fair equilibrium in which nobody ever deviates from proposing or accepting $(4,4)$. Treatment 3 is a hybrid equilibrium in which only player II proposes $(4,4)$.

enjoyed ample opportunity for learning. (The results look much the same with $\psi = 0$.) The remaining robots behave as described in the preceding section.

In Treatments 1 and 3, we took $f = (4,4)$ to reflect the fact that an attempt was made to condition the subjects on this outcome in the practice rounds. In Treatments 2 and 4, we took $f = (7.5, 2.25)$ and $f = (7, 2, 5)$ respectively to reflect the degree of success we enjoyed in seeking to condition the subjects on the outcome $(8,2)$. However, we would have done better by taking $f = (4,4)$ in all the treatments—as we do in the modified myopic best-reply model that we fit to the data in Binmore *et al* [11]. This observation is reflected in the fact that, although we have made no attempt to fit the current equilibrium model econometrically to the data, we do better by taking $\theta = 0.6$ in Treatments 1 and 3, and $\theta = 0.2$ in Treatment 2 and 4. Roughly speaking, this means that
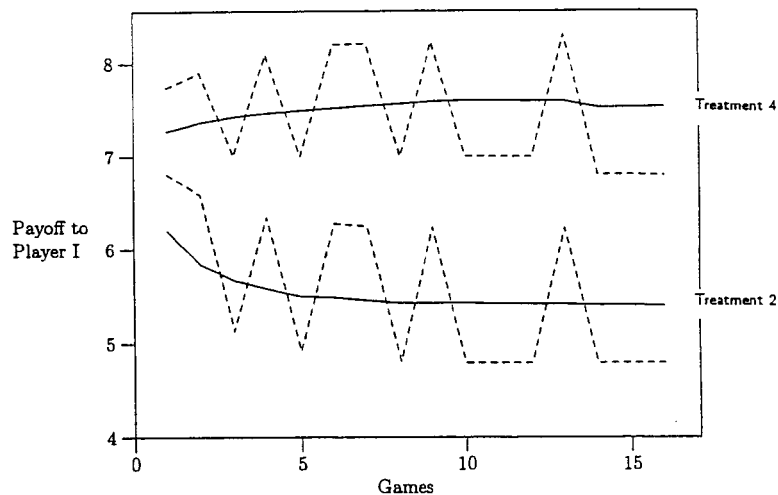
Figure 6: Equilibrium behavior in a perturbed Rubinstein model for Treatments 2 and 4. The unperturbed Rubinstein outcomes are $(4, 4)$ and $(8, 2)$ respectively. The parameters of the model are $\theta = 0.2$, $\phi = 0.5$, $\psi = 0.1$, with $f = (7.5, 2.25)$ in Treatment 2, and $f = (7, 2.5)$ in Treatment 4. Treatment 2 is a Rubinstein equilibrium. Treatment 4 begins as a hybrid equilibrium in which only player II proposes $f = (7, 2.5)$, but switches to a Rubinstein equilibrium when the fraction of robots becomes sufficiently small.

subjects are assumed to be more reluctant to abandon their conditioning when $f = (4, 4)$ than when $f = (7.5, 2.25)$ or $f = (7, 2, 5)$.

8. Conclusion

We have argued that anomalous data in bargaining experiments can often be explained without resorting to the extravagant claim that subjects act as optimizers with a large other-regarding component built into their utility function. We believe that a better explanation is that subjects are acting in accordance with a social norm which is adapted to a real-life game that differs from the game they are playing in the laboratory. When subjects fail to adapt their behavior to the laboratory game (with money payoffs) as in a minority of economic

experiments, we believe that the explanation is often to be found in the fact that the anomalous games have many Nash equilibria that are commonly overlooked. The field for such an explanation opens wider if one admits the possibility that the subjects may have a *small* other-regarding component built into their utility functions, or if there is some heterogeneity in the speed at which different subjects learn to adjust their behavior away from whatever social norm they brought with them into the laboratory.

In the main part of the paper, we explored the latter possibility using a perturbed version of the Rubinstein bargaining model with unequal discount rates. We find that the crude prediction of the unperturbed Rubinstein model must then be replaced by one of a rich variety of equilibria, some of which share the qualitative features of the available data.

Our general conclusion is that, before critics are entitled to argue that the income-maximizing hypothesis for experienced subjects should be abandoned in bargaining games or elsewhere, they first need to ask whether the behavior they observe is consistent with a neglected Nash equilibrium of the game or with a Nash equilibrium of some slightly perturbed version of the game.

References

[1] D. Abreu and F. Gul. Bargaining and reputation. *Econometrica*, 68:85–117, 2000.

[2] K. Binmore. Perfect equilibria in bargaining models. In K. Binmore and P. Dasgupta, editors, *Economics of Bargaining*. Cambridge University Press, Cambridge, 1987.

[3] K. Binmore. *Playing Fair: Game Theory and the Social Contract I*. MIT Press, Cambridge, MA, 1994.

[4] K. Binmore. *Natural Justice*. Oxford University Press, New York, 2005.

[5] K. Binmore. *Does Game Theory Work? The Bargaining Challenge*. MIT Press, Boston, 2006.

[6] K. Binmore, J. Gale, and L. Samuelson. Learning to be imperfect: The Ultimatum Game. *Games and Economic Behavior*, 8:56–90, 1995.

[7] K. Binmore, J. McCarthy, G. Ponti, L. Samuelson, and A. Shaked. A backward induction experiment. *Journal of Economic Theory*, 104:48–88, 2002.

[8] K. Binmore, P. Morgan, A. Shaked, and J. Sutton. Do people exploit their bargaining power? An experimental study. *Games and Economic Behavior*, 3:295–322, 1991.

[9] K. Binmore, A. Shaked, and J. Sutton. An outside option experiment. *Quarterly Journal of Economics*, 104:753–770, 1989.

[10] K. Binmore, J. Swierzbinski, S. Hsu, and C. Proulx. Focal points and bargaining. *International Journal of Game Theory*, 22:381–409, 1993.

[11] K. Binmore, J. Swierzbinski, and C. Tomlinson. An experimental test of Rubinstein's bargaining model. Else discussion paper, University College London, 2005.

[12] C. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, Princeton, NJ, 2003.

[13] E. Fehr and S. Gachter. Cooperation and punishment in public goods experiments. *American Economic Review*, 90:980–994, 2000.

[14] E. Fehr and S. Gachter. Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14:159–181, 2000.

[15] E. Fehr, S. Gachter, and G. Kirchsteiger. Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica*, 65:833–860, 1997.

[16] M. Kandori. Social norms and community enforcement. *Review of Economic Studies*, 59:63–80, 1992.

[17] D. Kreps, P. Milgrom, J. Roberts, and R. Wilson. Rational cooperation in the finitely repeated Prisoners' Dilemma. *Journal of Economic Theory*, 27:245–252, 1982.

[18] J. Ledyard. Public goods: A survey of experimental research. In J. Kagel and A. Roth, editors, *Handbook of Experimental Economics*. Princeton University Press, Princeton, 1995.

[19] A. Roth and I. Erev. Learning in extensive-form games: Experimental data and simple dynamic models in the medium term. *Games and Economic Behavior*, 8:164–212, 1995.

[20] A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50:97–109, 1982.

[21] R. Selten and R. Stocker. End behavior in finite sequences of prisoners' dilemma supergames: A learning theory approach. *Journal of Economic Behavior and Organization*, 7:47–70, 1986.

[22] J. Steiner. A trace of anger is enough: On the enforcement of social norms. CERGE-EI Working Paper, Prague, 2004.