



Bargaining and Dilemma Games: From Laboratory Data Towards Theoretical Synthesis

GARY E. BOLTON

Department of Management Science and Information Systems, 310 Beam, Penn State University, University Park, PA 16802, USA
email: geb3@psu.edu

Abstract

Bargaining and dilemma games have developed in experimental economics as fairly separate literatures. More than a few analysts are now persuaded that the patterns of behavior in these games are closely related, and considerable effort is being put into a search for models that bridge the gap between the two types of games. I focus on a handful of models that, when taken together, outline the conceptual issues, and provide a sense of the progress that has already been made.

Keywords: bargaining, dilemma games, bounded rationality, learning, motivation

JEL Classification: C78, C92, H41

But history shows that logic launched from introspection alone lacks thrust, can travel only so far, and usually heads in the wrong direction. —Edward O. Wilson

1. Introduction: Cultivating theory in a laboratory hothouse

The Handbook of Experimental Economics illustrates how things stood for quite some time. One chapter surveys bargaining games (Roth, 1995) while another surveys public goods games, the type of dilemma game most exhaustively studied in the lab (Ledyard, 1995). Both are marvelously insightful surveys. But if you crosscheck the reference sections, you will find little redundancy. For a long time, experimentalists treated bargaining and dilemma games like apples and oranges. That I am comfortable writing about bargaining and dilemma games together is indicative of an exciting project, just now taking shape in experimental economics: The search for models that succinctly capture what experimentalists have observed in their labs, models that span broad classes of games.

Experimentalists study a variety of bargaining games, ranging from two-person alternating offer bargaining (e.g., Binmore et al., 1985) to multilateral negotiation of coalition formation (e.g., Kahan and Rapoport, 1974). The feature common to bargaining games is of course the division problem; bargainers can individually benefit from a collective gain, but only if they reach mutual agreement on how to divide it. Experimentalists study a wide range of dilemma games, including the prisoner's dilemma (e.g., Selten and Stoecker, 1986),

public goods (e.g., Isaac and Walker, 1988), common pool resource (e.g., Budescu et al., 1995), investment (e.g., Berg et al., 1995), peasant-dictator (e.g., Van Huyck et al., 1995), and gift exchange (e.g., Fehr et al., 1993) games. All dilemma games feature a situation in which achieving efficiency requires a collective action that each participant has a financial disincentive to perform.

Lab data for bargaining and dilemma games does not square well with classical, full rationality game theory. Beyond this, the right interpretation has been a matter of some debate. For bargaining, the debate has centered on the role of fairness and the nature of strategic reasoning. For dilemma games, the debate has involved the relative weights that should be given to strategic reputation building, altruism, and reciprocity. The precise relationship between concepts like fairness, altruism and reciprocity has always been a bit unclear. I think this is one reason that bargaining and dilemma games have tended to end up in different conference sessions (with the lack of cross pollination that this implies). Another reason is that bargaining and dilemma games are often, from the perspective of classical game theory, strategically distinct. Many dilemma games exhibit a dominant strategy equilibrium, while bargaining games are typically examined through the lens of subgame perfection.

Whatever the reasons, things are changing. More than a few analysts are now persuaded that the patterns of behavior in bargaining and dilemma games are closely related, and considerable effort is being put into constructing models that bridge the gap between the two types of games. While it is far too early to say what the final product of this line of research will be, I believe that the major theoretical components, or at least their prototypes, are on the table. Some of the components can be gleaned from early models that confine themselves to explaining one or the other type of game. A few come from recent, explicit attempts at building a bridge.

The effort is in its nascence. As such, this paper should be viewed as a new product sketch, not as an attempt at a final blueprint. I present a framework for organizing the various component parts, and I sample the modeling work being done on each. The paper is not a comprehensive survey. I focus on a handful of models that, when taken together, outline the conceptual issues, and also provide a sense of the progress that has already been made. I hope the reader gains some insight into how the models, or at least the concepts they deal in, will eventually fit together. I close with my own thoughts on the matter.

This brings me to the question of what we are hoping to accomplish by building models from lab data (as opposed to confining ourselves to testing models that originate from outside the lab). The idea, in essence, is to use the lab as a hothouse for cultivating theory: Existing lab data guides model construction. Once constructed, the model is subjected to a new round of tests. Depending on the results, the model is either refined or abandoned. This sort of feedback loop plays to the lab's comparative advantage: control. The lab environment can be customized to the model, producing data that is subject to a relatively low level of confounding. Data can be produced at will (assuming funding, of course). In sum, laboratory control makes for a relatively short and informative feedback loop, something that should facilitate rapid model development. The models that survive this process are then ready to be tested outside the hothouse, in the field, the ultimate arbiter of validity. The models most likely to survive are those that are most robust. Hence the importance of bridging gaps—like the one between bargaining and dilemma games.

2. Some simple, and some still simpler, games

When theorists find a problem is too complex, they pare it down to a simpler, more easily understood case, conquer this, and then return with what they have learned to the more complex case. Beyond practicality, this research strategy embraces parsimony: Find a simple explanation that fits the simple case, and run with it as far as you can. The same approach can be applied in experimental design, to help untangle the behavior we observe in experiments. In the lab, the approach translates to paring down a game to its essence, getting a reasonably clean interpretation of behavior in the simpler environment, and then testing this interpretation out on the more complex cases. If we are successful, we have one explanation for a wide variety of phenomena. With this in mind let's look at some bargaining and dilemma games to get an idea of the issues.

Figure 1 displays cooperation rates for the partners condition of a finitely repeated prisoner's dilemma experiment performed by Andreoni and Miller (1993). Each game matched two players for 10 rounds of a standard prisoner's dilemma.¹ The 10-round length was common knowledge to the players. We can see from the figure that cooperation rates are high in the early rounds, and then tail off at the end. Selten and Stoecker (1986) also study a finitely repeated prisoner's dilemma experiment, and report similar results.

There is a pattern to the individual behavior behind figure 1. Players tend to cooperate early on. At some point, one player defects, after which neither player cooperates further. Few players cooperate until the very end. Results like these seem to imply some sort of 'strategic' reciprocal behavior, although the classical strategic solution for this game cannot sustain any cooperation, reciprocal or otherwise. The one-shot prisoner's dilemma has a dominant strategy (defect). When the game is iterated for a known, finite number of rounds, subgame perfect equilibrium predicts that no cooperation is possible because of the unraveling problem. Roughly speaking, given that defecting is a dominant strategy in the single-shot game, one should cooperate only if doing so encourages ones partner to cooperate in future rounds. But then neither party should cooperate in the last round because no future cooperation is possible. But then there is no incentive to cooperate in the next-to-last round, etc. Hence the type of reciprocity on display in figure 1—assuming it is in fact some sort of reciprocity—must fall outside the realm of classical game theory.²

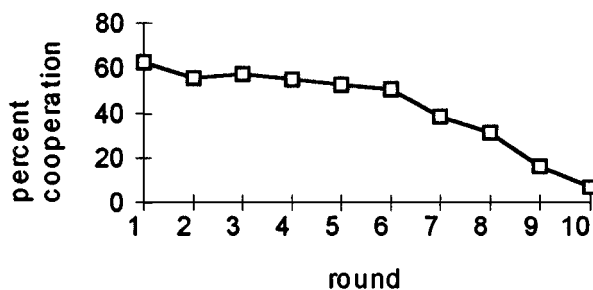


Figure 1. Mean cooperation rates in a 10-round prisoner's dilemma, partners condition (Andreoni and Miller, 1993).

Most of us think that human behavior involves strong elements of rationality. And the fact that just about all players leave a repeated prisoner's dilemma experiment with substantially more money than the full rationality theory would predict should give pause to anyone who wants to write this behavior off as totally irrational. If we think there is an element of truth in the full rationality model, then an important question becomes: What tenets of full rationality are being violated by the behavior we observe? The right answer says a great deal about how to go about constructing a more accurate model.

In fact, Andreoni and Miller ran their experiment to test the well known Kreps et al. (1982) model, which alters the classical model by a single precept. Instead of supposing that it is common knowledge that all players are rational, suppose there is a chance that the person you are playing is altruistic, in the sense that he wants to cooperate, but only if you do so. Kreps et al. show that even if this chance is quite small, it pays to invest in reputation building; that is, it pays to cooperate for awhile before defecting, much as we see in figure 1. Two conditions not shown in the figure manipulated the belief of an altruistic opponent through the use of computer partners. The observed behavior was consistent with what we would expect from the Kreps et al. model.

There was also a fourth condition, however, called strangers, in which partners were randomly paired for one-shot prisoner dilemma games. In these games there is no incentive for reputation building, and so by the Kreps et al. model, we would expect to see no cooperation. But cooperation was nevertheless observed. The amount, while less than that displayed in figure 1, was substantial and consistent.

Andreoni and Miller interpret their experiment as evidence that reputation building is a part of what we are seeing in figure 1. But they also interpret the results from the strangers condition as evidence that, rather than simply believing that some players are altruistic, many players *are* altruistic.

By way of explaining their own experiment, Selten and Stoecker throw a third explanation into the mix, one that the data from Andreoni and Miller's experiment does not entirely rule out. Selten and Stoecker argue that cooperation does not unravel because there is a learning aspect to behavior that subgame perfection fails to capture. Given that there is typically no further cooperation after one player defects, the optimal round for first time defection is the one just prior to the round your partner is planning for first time defection (at least if your goal is to make the most money possible). Selten and Stoecker propose a learning model in which players tend to change their first defection round in the direction that would have been better in the last game they played. They then show that the model tracks their data.

The pattern of cooperation observed in Andreoni and Miller's strangers condition remained stable after 200 iterations, which could be taken as evidence against a learning explanation. But it could also indicate that learning is very slow, or that learning converges to something other than the subgame perfect equilibrium.

When we begin thinking down these lines, Andreoni and Miller's experiment, which is in some ways quite simple, begins to look quite complex. The difficulty is that it is not immediately apparent how to pull the three hypotheses—reputation building, altruism, and learning—apart. The beliefs that are central to the reputation building model are not directly observable. Altruism can be confounded with strategic considerations like

learning. Learning takes many shapes, and not all are well understood. Can we separate these hypotheses in the context of the finitely repeated prisoner's dilemma? That is, can we say which hypotheses are valid, and if more than one is valid, can we say how they fit together? In the context of the finitely repeated prisoner's dilemma, the difficulties are formidable.

Can we learn something by paring the game down? I think that we can. The dictator game is a radical simplification that is nevertheless, in a strategic sense, closely related to the one-shot prisoner's dilemma. In the dictator game, the 'dictator' unilaterally divides a sum of money between self and another, the recipient. The game is so simple that it is not really a game—it's a one-person decision problem. But it is also perhaps the simplest possible example of a dominant strategy: The dictator should keep all of the money.

Beginning with Kahneman et al. (1986), the dictator game has been the subject of many experiments. Dictators regularly fail to keep all of the money. Dictator giving for the Forsythe et al. (1994) experiment is reproduced in figure 2 (later on, we will get to the ultimatum game data in the figure). Each dictator played exactly once. I know of few experiments in which subjects sequentially play multiple dictator games, no doubt because the game is so transparent, with no strategic principle to learn.³ In fact, this is precisely why I find the game so illuminating: People deviate from the dominant strategy in a situation where there is no easy role for learning.

Nor is there any easy role for reputation building, since subjects play with one another only once, and as in the Forsythe et al. experiment, they usually play anonymously. It has been conjectured that dictators give to build a reputation with the experimenter. But experiments designed to test this conjecture report substantial giving even when the experimenter cannot associate particular gifts with particular subjects.⁴ In fact, evidence on whether the hypothesis can explain any of the gift giving is mixed. Roth (1995) summarizes much of the research, and discusses an alternative interpretation for the positive evidence having to do with the directions given to subjects.

This is not to say that learning and reputation building cannot play a role in the finitely repeated prisoner's dilemma. They might well play a role; the dictator game results do not

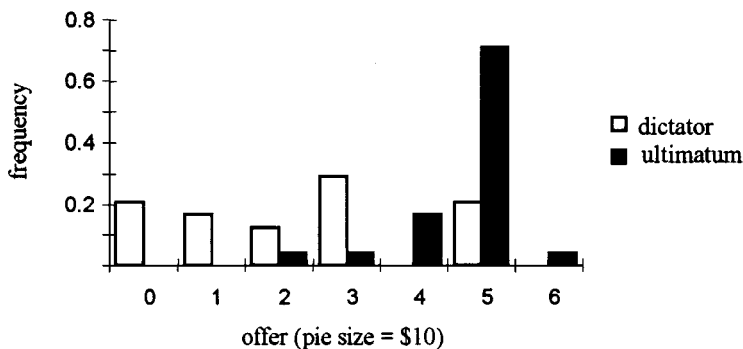


Figure 2. Amounts offered to the recipient in dictator & ultimatum games (Forsythe et al., 1994).

speak to this issue. What the dictator results do suggest is that even when we strip learning and reputation building issues away, some factor remains, call it the dictator factor, that drives many people to deviate from the dominant strategy. It seems sensible then to try to understand the dictator factor in isolation, before attempting to add learning or reputation building to the mix, both of which may interact with the dictator factor in what are, without further knowledge of the three separate influences, difficult to predict ways.

Is the dictator factor altruism? Let's come back to that later.

The dictator game does something else important. It provides a rudimentary link between dilemma and bargaining games. Dominant strategy is the link to dilemma games. The dictator game can be transformed into a bargaining encounter by adding a single action to the strategy space: Allow the recipient to reject the proposal and end the game with both players receiving nothing. This is the ultimatum game, first studied by Güth et al. (1982), and it is perhaps the simplest possible bargaining game. The subgame perfect equilibrium for this game has the second mover (formerly the recipient) accepting any positive offer, and as a consequence, the first mover (formerly the dictator) offering no more than the minimum positive amount possible. But when the ultimatum game is played in the lab, we see another well-known anomaly: the second mover sometimes rejects a proposal, even when doing so leaves her with less money than would accepting.

The results from dictator and ultimatum games are quite stable. While some learning is evident as ultimatum bargainers gain experience, what bargainers learn does not approach the subgame perfect equilibrium. After some rounds (10 or so), play settles down with a typical first mover offering around 40%, and with substantially smaller offers persistently rejected. Roth (1995) reviews the substantial evidence for ultimatum game robustness with respect to a variety of factors. Forsythe et al. show that dictator giving is stable with respect to time. Hoffman et al. (1994) replicate the Forsythe et al. distribution of giving using the Forsythe et al. instructions. Bolton et al. (1998) demonstrate that the amount the dictator gives is stable with respect to various game manipulations. Dictator giving is sensitive to framing effects, although no more so than other dilemma games.⁵ Giving behavior is not restricted to people: Capuchin monkeys and chimpanzees give food in what is an animal version of the dictator game (de Waal, 1996, chapter 4).

One might be tempted to conclude that there is no role for game theory here; that is, one might conclude that there is no meaningful strategic factor at work in these simple games. Forsythe et al.'s experiment demonstrates that this is wrong. The offers in figure 2 are significantly higher in the ultimatum game than in the dictator game. It is hard to escape a strategic explanation for why this would be so: If first movers did not concern themselves with the behavior of the second mover, they should do as dictators do in the dictator game, where the second mover's behavior is definitely not an issue. On the other hand, the higher ultimatum offers are precisely what we would expect if first movers are strategically minded, and if they anticipate the tendency of second movers to reject offers deemed too low (rejected offers lead to a very poor payoff for both players).

It appears then that a successful model need include a strategic component. But it also appears it need include something non-strategic, or possibly something that is strategic in a way that falls outside the classical game theory notion of strategic thinking. The models I describe can be classified by how they handle this conundrum.

3. A taxonomy of models

One of the great strengths of experimental economics is its custom of tolerating—actually inviting—radically different approaches to open questions. The models I describe here reflect this sort of diversity. Bounded rationality unites them—although only if we agree to interpret the term ‘bounded rationality’ very liberally, to include all models outside of the classical, full rationality approach. More substantially, all of the models are strategic in nature, and so all satisfy the base criteria set down by the Forsythe et al. experiment.

While classical, full rationality game theory does not provide an accurate description of behavior, it does provide a useful benchmark for classifying the various theories. It is easiest to see what I have in mind through an example. Consider a standard, single-shot prisoner’s dilemma in which each player must simultaneously choose to either cooperate or defect. The full rationality argument for playing the dominant strategy can be divided into three components:

Motive: A player prefers having more money to having less.

Choice: A player takes the action that she believes best satisfies Motive.

Cognition: A player can identify the action that best satisfies Motive.

The dominant strategy argument is assembled as follows: Each player sees that defecting maximizes the amount of money she will make independent of what the other player does (motive plus cognition). Each player then chooses defect (choice).

Most bounded rationality theories dealing with bargaining or dilemma games depart from full rationality on some subset of the three components (motive, choice, cognition) and rely on full rationality arguments, sometimes implicitly, for the other components. Below, I classify each model by major component of deviation. The taxonomy is mostly for purposes of convenience; in some cases, a model’s classification is a matter of judgement.⁶

I create a fourth category for *evolutionary* models. These are directed at a somewhat different question than the other models: Evolutionary models offer an answer to *why* rationality is bounded, whereas the other models are focused on *how*. The principle focus of the present paper is models that fall into the how-category, but I also want to mention some evolutionary models, if only briefly, because I think the how-models raise why-questions too important to be ignored.

For each category, I highlight a handful of models that illustrate some basic issues. The assessments expressed are my own, and the reader may well find that he or she disagrees. As I mentioned before, what we are doing here is a new product sketch. Disagreement and debate are an inevitable part of development.

4. Models of motivation: Enriching narrow self-interest

Some analysts (I am one) interpret the results of the Forsythe et al. experiment as a challenge to the classical theory’s assumption concerning what motivates people. More specifically, the fact that dictators in the dictator game give money, and second movers in the ultimatum game reject money, challenges what I call the *narrow self-interest* assumption, that “more

money for myself is always preferred to less.” I use the ‘narrow’ qualifier because it is not clear that “more money for myself is always preferred to less” is a sufficient description of self-interest, particularly from the perspective of biology, a point dealt with explicitly by the evolutionary models.

None of the motivation models discussed here completely abandons narrow self-interest. Instead, each posits an additional motive to interact with narrow self-interest. In all cases, the motive is added as an argument to some form of preference function. Models differ on what is added, or if they agree on this, they differ on the nature of the interaction.

4.1. Altruism

One approach is to suppose that individuals have some tendency towards altruism. The usual technique is to add a variable to the utility function representing other’s consumption.⁷ Individuals then distribute their income in accordance to the marginal rate of substitution between own and other’s consumption, factoring in any strategic considerations. This approach has a long history in the political economy literature where it is used to explain the production of public goods (Hochman and Rodgers, 1969; Becker, 1974).

Restricting attention to two people, i and j , we would typically work altruism into i ’s utility function by writing

$$U_i = U_i(y_i, y_j) \tag{4.1.1}$$

where y_k represents person k ’s ‘income’, and U_i increases in both arguments. This formulation can obviously describe the results from the dictator game. But for it to be a meaningful explanation, preferences for giving must exhibit some stability. I say ‘some’ because I do not think it is reasonable to demand that preferences over distribution be completely fixed and immutable, anymore than it is reasonable to demand this of preferences over clothing styles or public policy, things that economists put into utility functions all the time. In fact, as noted in Section 2, there is evidence that giving in the dictator game is stable, at least up to framing (a factor widely suspected of influencing preferences for clothing styles and public policy).

Public goods models that incorporate the specification of altruism given in (4.1.1) have a neutrality property that runs counter to field observation. Specifically, the models imply that government donation to the public good, funded by lump-sum taxes, crowds out private donation dollar-for-dollar. The argument can be stated succinctly in terms of (4.1.1): Imagine that i voluntarily contributes some amount z to j ’s income (the public good in this example). Suppose that i is the only contributor. Observe from (4.1.1) that i ’s preferences are completely determined by the final allocation of income between i and j . Person i should therefore be indifferent to whether he contributes z voluntarily or involuntarily through a tax transfer. Hence, if a law is passed to force i to contribute $x \leq z$, i ’s voluntary contribution should fall by exactly x . This one-to-one crowding out, or neutrality property, is also a property of Nash equilibrium for public good games where more than one person contributes. In fact, the neutrality property is quite robust, and even holds for certain distortionary taxes (Bernheim, 1986).

Field studies of charitable giving, however, report that actual crowding out is quite small, with estimates ranging from 5 to 28%. One explanation for the inconsistency, advanced by Andreoni (1989), is known as the impure altruism hypothesis. Andreoni argues that people receive some value—a “warm glow”—from the act of voluntary gift giving. Public transfer is an imperfect substitute, and this makes for incomplete crowding out. Andreoni (1989) introduces a warm glow parameter as an additional argument into altruistic utility functions, and shows that the resulting model is consistent with a variety of field phenomena related to crowding out. There is also some laboratory evidence for warm glow in public goods (Andreoni, 1993), and dictator (Bolton and Katok, 1998) experiments, although the crowding out observed, about 75% in both cases, is much closer to complete than what is observed in the field.

Other laboratory evidence that altruism, or at least impure altruism, plays a role in public goods games is reported by Andreoni (1995a), and Palfrey and Prisbrey (1997). They differ substantially on the estimated amount of altruism, although both agree that other factors are at work as well. More on these other factors in Section 6.1.

Experiments turn up another interesting phenomenon that altruism has a difficult time accounting for. Bolton et al. (1998) found that the proportion of a fixed amount of money given away by the dictator is independent of the number of recipients. Selten and Ockenfels (1998) replicated this ‘fixed sacrifice effect’ in the solidarity game, a kind of dictator game in which the number of recipients depends on a roll of a die. They demonstrate that this effect is not easily reconciled with an altruistic utility formulation like (4.1.1).

Warm glow and fixed sacrifice effects imply that the formulation in (4.1.1) is at best a rough approximation of actual giving behavior. Still, we might, on the grounds of parsimony, be tempted to brush these effects aside, and confine ourselves to (4.1.1) as a base for explaining dictator games in particular, and hopefully, dilemma games in general. But in terms of the broader context we have set out for ourselves, (4.1.1) is incontrovertibly inadequate: a utility function that is increasing in both own and other’s income has no chance of explaining why ultimatum game second movers turn down money in favor of both bargainers receiving nothing.

4.2. *Distribution*

The comparative model (Bolton, 1991) posits motives consistent with ultimatum second-mover behavior. According to this model, people are motivated by their own pecuniary payoff as well as by a relative payoff, a measure of how the pecuniary payoff compares among players. Ultimatum game second movers turn down money—sacrifice own pecuniary payoff—in order to gain relative payoff (all receiving nothing is an equal split). Beyond ultimatum games, the comparative model is consistent with a variety of phenomena observed in laboratory play of two-period alternating offer bargaining games.

But the comparative model fails absolutely with the dictator game. And it is then perhaps not surprising that the model is of little use in explaining dilemma games of any type. The reason for the dictator game failure is straightforward: Comparative model utility functions are non-decreasing in both own pecuniary and relative payoffs. Consequently, no one should be inclined to give a gift to another.

The ERC model (Bolton and Ockenfels, 1997) amends the interaction between pecuniary and relative arguments in a way that directly accounts for the dictator game, but preserves the ability to account for the ultimatum game. ‘ERC’ stands for *e*quity, *r*eciprocity and *c*ompetition. Bolton and Ockenfels (1997) show that the model can explain phenomena observed in several bargaining (equity) and dilemma (reciprocity) games, as well as in some simple market contests (competition). Bolton and Ockenfels (1998) show that ERC is consistent, to a fair degree of detail, with the three-person bargaining experiment reported by Güth and van Damme (1998).

Bolton and Ockenfels characterize preferences in terms of a motivation function, which may be thought of as a special class of expected utility functions. Use of the term ‘motivation function’ emphasizes that the absolute-relative trade-off may not be immutable, although it need be stable for the duration of the experiment. In an n -player game, player i maximizes the expected value of his or her motivation function, $v_i(y_i, \lambda_i)$, where λ_i is i 's relative payoff, and

$$v_i(y_i, c, n) = \begin{cases} \frac{y_i/c}{1/n} = \frac{n}{c}y_i, & \text{if } c > 0 \\ 1, & \text{if } c = 0 \end{cases}$$

where $c = \sum_{j=1}^n y_j$ is the size of the pie that is distributed among all players.

The motivation function is then characterized by several axioms. For present purposes, it suffices to display a typical function in a picture. Towards that end, write $k = c/n$, the average pecuniary payoff. Then $k\lambda_i = y_i$ and we can write $v_i(y_i, \lambda_i)$ as $v_i(k\lambda_i, \lambda_i)$. Holding k fixed, a typical motivation function looks like figure 3. The peak of the function at $\lambda_i = 1.5$ indicates that, as a dictator in the dictator game, i keeps 1.5 times the average payoff, k , for himself, and gives the rest to the other player. The function crossing the horizontal axis at 0.38 indicates that, as a second mover in the ultimatum game, i is indifferent between

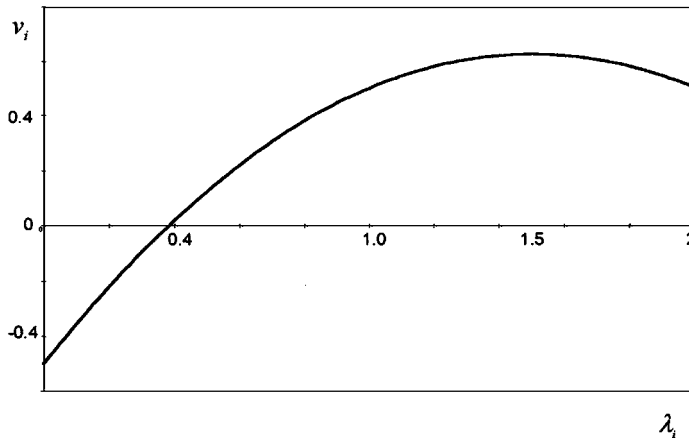


Figure 3. An example motivation function.

accepting or rejecting $0.38k$. The model accommodates different people with different dictator and ultimatum cutoffs, thereby accounting for the heterogeneity of behavior on display in figure 2.

These preferences are conceptually distinct from the altruism formulation of (4.1.1). We usually think of altruism in terms of concern for the welfare of other individuals, and hence altruism is usually associated with a concern for how payoffs are distributed across individual members of the group. But in the ERC model, players are assumed to care only about the division between self and the rest of the group, not about the payoff distribution across all players. As a result, the model is consistent with the fixed sacrifice effect described by Selten and Ockenfels. ERC does not, however, square with the warm glow effect. In fact, ERC preferences have the same neutrality property as altruistic preferences (4.1.1).

A model by Fehr and Schmidt (1997) makes similar predictions to those of ERC, and for many of the same bargaining, dilemma, and market games. Like ERC, the Fehr-Schmidt model incorporates an equity term along with an own earnings term into the utility function. There are two differences. One is that, in the Fehr-Schmidt model, inequality between self and all others matters instead of inequality between self and the average earnings of others. Second, most (although not all) of the results in the Fehr-Schmidt model suppose complete information, whereas those for ERC suppose that individual characteristics of motivation functions are private information.

Altruism, ERC and the Fehr-Schmidt model all fail to capture the warm glow effect because they focus exclusively on distribution. Warm glow implies a concern beyond distribution.

4.3. *Intentions*

While intention models do not explain the warm glow effect, they do go beyond distribution. These models are sometimes referred to as ‘psychological games’ (Geanakoplos et al., 1989). The model of this type most often cited by experimentalists is Rabin (1993). It begins with a utility function that reflects the notion that people like to help those who help them, and hurt those who hurt them. Hence the emphasis is on intentions: A player cares whether another player’s actions were intended to help or hurt. Rabin shows that the predictions of the model are qualitatively consistent with what we observe in certain games, including the one-shot prisoner’s dilemma.

Rabin’s model technically applies only to games in the normal form.⁸ The intuition behind Rabin’s model, however, is easy enough to apply to a simple extensive form game like the ultimatum game: a second mover rejects a small offer to punish the first mover for what the second mover interprets as a hurtful act. This is a more elaborate explanation for rejections than that offered by distribution models. According to the intention explanation, the second mover must first decide whether the offer was intended as a hurtful act; if so, he must then determine whether rejecting is a fitting punishment. He must consider the tradeoffs—compare the payoff distributions—that result from accepting or rejecting. According to the distribution explanation, the second mover simply compares the payoff distributions.

There is some lab evidence that beliefs about intentions play a role in simple games, but there is also evidence that distribution, in and of itself, is a very substantial factor. Intentions, for example, cannot explain giving in the dictator game, since the recipients have no opportunity to do anything of any sort, either helpful or hurtful. Experiments on the ultimatum game find evidence that beliefs about intentions are responsible for some, but not all, rejections (Blount, 1995; Kagel et al., 1996). Charness (1996) finds some evidence for intentions in the gift exchange game, but he also finds a very considerable role for distribution independent of intentions. Bolton et al. (1998) study a simple dilemma game that allows both helping and hurting behavior, and they find that distribution alone is sufficient to explain their data.

To me, the issue here is to what extent it is necessary to account for intentions in order to *predict* helping or hurting behavior. Introspection suggests to many of us that intentions are crucial to the cognitive process of reciprocating, and for this reason it may be difficult to imagine that they are not decisive to the outcome. But introspection can be misleading. For instance, introspection suggests to most of us that rational decision making requires strict emotional control. But brain science has demonstrated to the contrary, that emotions play a crucial role in the proper functioning of rational decision making; without emotional input, rational thought disintegrates.⁹ Something similar may be at work with intentions. A person may well think in terms of rewarding the kindness of another, but he may nevertheless assign the same, or approximately the same, payoff distribution as in a situation where intentions are irrelevant.

5. Models of choice: Satisficing and rules-of-thumb

Modifying the choice rule is what is traditionally meant by bounded rationality. These models move away from the maximization assumption, and towards satisficing or rules-of-thumb. Choice models differ from one another with respect to how far they move away from formal optimization, as well as with respect to the alternative decision rule considered.

5.1. *The theory of equal division payoff bounds*

This is a model put forward by Selten (1987) to explain coalition bargaining games in characteristic form. The model is unique among those I review in that it is based on cooperative game theory. (Selten, 1987, reviews another prominent cooperative game theoretic model aimed at experiments, the bargaining set theory of Aumann and Maschler, 1964). In the theory of equal division payoff bounds, players make decisions about coalition payoffs based on aspiration levels. The latter are constructed from considerations of ordinal power comparisons, as well as equitable division. The model is an ‘area theory,’ meaning it predicts a range of outcomes without attempting to say how likely each outcome is. Predicting a range rather than a point is in an important sense a strength of the model: Coalition bargaining exhibits substantial heterogeneity with respect to outcomes. So while a model that makes a point prediction is more precise than an area theory, it will also be quite limited in terms of potential accuracy.

Area theories are also more complicated when it comes to comparisons with other theories. A theory that allows four different outcomes will trivially be at least as accurate, in terms of hit rate, as a theory that predicts a strict subset of these outcomes. Broader is not necessarily better. To remedy this difficulty, Selten and Krischker (1982) propose a measure of predictive success that is, roughly, the relative frequency of correct predictions minus the random frequency of correct predictions. Selten (1987) uses this measure to demonstrate that the theory of equal division payoff bounds fits the data better than a bargaining set model for a fairly large collection of coalition experiments.

Kuon and Uhlich (1993) provide experimental evidence for another area theory, the negotiation agreement area, as it applies to a two-person characteristic function game.

5.2. *Reciprocal obligation*

Sugden (1984) proposes a model of public good contribution based on the idea that people feel a moral obligation to contribute something if everyone else does so as well. This is a non-cooperative model in which a person maximizes utility subject to a moral obligation constraint. Sugden writes i 's utility function as

$$U_i = U_i(q_i, z)$$

where q_i is i 's contribution to the public good, and z is the quantity of the public good provided (z is a function of the vector of player contributions). U_i decreases in q_i and increases in z . Sugden then lays down what he called the 'reciprocity principle,' which I state somewhat roughly: Let q_i^G be the contribution that maximizes U_i under the assumption that everyone contributes the same amount. The reciprocity principle then says that i 's obligation is to contribute at least q_i^G if everyone else does so, or to contribute at least as much as everyone else if someone other than i contributes less than q_i^G . An equilibrium vector of contributions is one in which each player's contribution is the smallest one that meets his or her obligation.

Sugden demonstrates that his model fits various stylized facts culled from field and lab. One interesting implication of the model is that an increase in one person's contribution should tend to induce others to increase their own contribution. Hence Sugden's model advances moral obligation as the reason why charitable giving is not completely crowded out by outside giving. It is interesting to think about how Sugden's moral obligation explanation differs from Andreoni's warm glow explanation (something I return to at the end of this section).

The spirit of the model is also consistent with the fixed sacrifice effect (Section 4.1). Following a rule-of-thumb based on moral obligation to share some fixed portion of one's earnings with others strikes me as a very plausible explanation for why giving is insensitive to the number of recipients. Taking it a bit further, we might conceive of rejections in the ultimatum game as the second mover punishing the first mover for failing to meet his moral obligation.

Brandts and Schram's (1996) theory of cooperative gain seeking is, from a technical perspective, quite different from Sugden's model, but nevertheless similar in flavor.

Cooperative gain seekers are individuals who do contribute to a public good when they expect that a cooperative gain is possible, but do not contribute otherwise. Isaac et al. (1994) also develop a model based on gains from cooperation. They attempt to link the level of contribution to the public good game parameters that are known to robustly effect contribution levels. (Also see Ledyard's, 1995, excellent summary of the factors known to influence the level of public goods contribution.)

5.3. *Focal points*

Schelling (1960) introduced the concept of 'focal point' to describe allocations or division points that are for cultural or other reasons prominent in peoples' minds. Focal points tend to end up as game outcomes. One of the clearest examples of the focal point phenomenon can be found in the binary lottery bargaining experiment of Roth et al. (1981). In this study, bargaining pairs negotiated over the division of 100 lottery tickets, each ticket providing a chance at a prize. The value of the prize differed across bargainers. When prize values were commonly known, observed bargaining settlements tended to cluster around one of two focal points: equal ticket division or equal expected lottery value.

Roth (1985) takes a step in the direction of formally modeling the influence of focal points in binary lottery bargaining games. He considers the case when chip values are commonly known to both bargainers, and supposes that bargainers restrict their offers to one of the two focal points. The model is a normal form bimatrix game in which proposals are made simultaneously, and agreement is reached only if the proposals are compatible. The game has a mixed strategy equilibrium that agrees fairly well with the proportion of disagreements previously observed in experiments.

In what way do choice models differ from motivation models? I think there are two ways of looking at this question. Viewed one way, the difference is substantive (testable). Viewed the other way, the difference is one of research approach.

The substantive difference is clear from a comparison of formal structure. Motivation models incorporate the thrust of their arguments into preference functions, while choice models either place restrictions on the optimization of preferences, or substitute satisficing for optimization. As such, motivation models treat equity and moral obligation as goals people pursue, while choice models treat these same objects as rules-of-thumb, tools used in pursuit of a goal. On this score, I think that dictator giving in dictator games, and second mover rejections in the ultimatum game cut decisively in favor of the motivation models. For me, these actions are too clear-cut to be attributed to satisficing tools. I do not rule out the possibility that in more complicated games, principles of fairness and morality play a rule-of-thumb role, but the simple games imply there is more to it than this.

To see the research approach difference think about the models in intuitive terms. Sugden's model attributes the failure of total crowding out to moral obligation restricting optimization, while Andreoni says it results from a warm glow parameter in the utility function. But can't a person get a warm glow from the feeling of having done their moral duty? Likewise, Selten's cooperative coalition model portrays bargaining as a tug-of-war between negotiating power and equity principles, while the non-cooperative ERC model portrays it as an interaction between strategic opportunity and concern for distribution.

Thought of this way, motivation and choice models are trying to get at very similar things, the difference being research approach: Motivation models make fairly general statements about the motives driving behavior, statements like ‘people care about distribution’, or ‘intentions to help or hurt matter.’ Choice models make more specialized statements. Selten’s equal division payoff bounds, Sugden’s reciprocity principle, Roth’s binary lottery focal points are all very specific concepts pertaining to specific games.

Motivation models are stated in a way that is intended to make them applicable to a relatively large domain. Probably too large. They will have to be pared back as more data becomes available. Choice models are engineered for very specific situations. It remains to be seen by how much they can be generalized. Hence motivation and choice start at opposite ends, but may well end up in the same place.

The behavioral concepts choice models use—equity, obligation and focality—are in principle, quite general. Bacharach and Bernasconi (1997) take an interesting step towards a general model of focal points (although they do not address the types of games considered here). The basic idea is that perceptual clues serve to frame the game, and the frame in turn determines what equilibrium will be selected for play. The paper includes the results of some experimental tests.

6. Models of cognition: Reputation building and errors, learning

Whereas choice models posit that people willfully follow rules-of-thumb, cognitive models characterize behavior in terms of either errors, or in terms of a learning process in which the rule that governs the choice changes with experience.

6.1. Reputation building and errors

McKelvey and Palfrey (1992) construct a model to explain an experiment they ran on the centipede game, a type of dilemma, in which two players alternate opportunities to either ‘take’ or ‘pass.’ If a player takes, the game ends and the taking player gets the larger share of the payoff pie. If instead the player passes, the game advances to the next round where the payoff pie doubles. A passing player will gain financially (relative to having taken) only if his playing partner passes in the next round, and strictly loses if his playing partner takes. McKelvey and Palfrey study both four and six round versions of the game. In both cases, all Nash equilibria, and consequently all Nash refinements, predict that the first mover will take in round 1.

But McKelvey and Palfrey find that only 37 of 662 games end with the first mover taking in round 1. At the other extreme, 23 of the games end with both players having passed at every opportunity. The vast majority of games are somewhere in between.

McKelvey and Palfrey go on to construct an incomplete information model based on reputation building, similar to Kreps et al.’s approach to the finitely repeated prisoner’s dilemma. In the McKelvey and Palfrey model, players attach some small probability to the likelihood that their playing partner is an altruist, defined in this instance as someone who always passes. The model also allows for some proportion of the population to actually be altruists.

The model allows both decision error as well as belief error. Belief error refers to player mistakes in forecasting the proportion of altruism in the population. The subsequent data analysis indicates that both types of error play a significant role. McKelvey and Palfrey estimate the proportion of altruists in the population to be on the order of 5%. They find that beliefs about the proportion of altruists are, on average, consistent with this estimate, although there is substantial individual bias. They also find evidence that the amount of decision error decreases as players gain experience with the game, implying that players learn with experience.

McKelvey and Palfrey (1995) describe a more general model of error-prone decision making. Fudenberg and Levine (1997) present a model for measuring the size of the errors in experimental games. Palfrey and Prisbrey (1997) find evidence for decision-error, and a small amount of impure altruism in a public goods game. Anderson et al. (forthcoming) describe a model that allows both altruism and decision-error. They show that the model is consistent with several phenomena observed in public goods games. They estimate the model with data from Isaac and Walker (1988) and Isaac et al. (1994) and find evidence for both pure altruism and decision-error.

McKelvey and Palfrey's (1992) study provides evidence for reputation building in dilemma games where the same players interact repeatedly. The model provides for some altruism in the population, and in this sense there is an overlap with motivation models. The reader might find the 5% estimate of the proportion altruists at odds with the proportion that give in the dictator game (see figure 2). But there is no inconsistency here. McKelvey and Palfrey use a tighter definition of altruism than what is given in (4.1.1). Specifically, McKelvey and Palfrey define an altruist as somebody who puts sufficient weight on the payoff of his playing partner that he passes at every opportunity. By this definition, altruists are precisely those players for whom the game provides no incentive for strategic play. In contrast, motivation and choice models allow for players who put some weight on their playing partner's payoff but still have an incentive to behave strategically, clearly a broader group.

6.2. Reinforcement learning

Roth and Erev (1995) describe a simple reinforcement learning algorithm that predicts the dynamic path of play as people learn about both the game and the behavior of other players. They use their model to explain why behavior tends away from perfect equilibrium in ultimatum games as bargainers gain experience, but towards very similar looking perfect equilibria in two other types of games, including a public goods game. The key idea is that different payoff outcomes off-the-equilibrium path reinforce experimentation in different ways. The right reinforcement can lead play to the subgame perfect equilibrium, while other types of reinforcement lead play away from perfect equilibrium, perhaps indefinitely.

The basic mechanics of the Roth-Erev model are as follows: Each player n begins the first round of play, $t = 1$, with an initial propensity to play his k th pure strategy, given by some number $q_{nk}(1)$. Repeated play modifies these propensities through a process of adaptation. If player n plays his k th pure strategy in round t , and receives the payoff x , then the propensity to play k is updated to $q_{nk}(t + 1) = q_{nk}(t) + x$. The probability that k

gets played in round t is $q_{nk}(t) / \sum q_{nj}(t)$ where the summation is taken over all of n 's pure strategies j . Roth and Erev consider a variety of modifications on the basic mechanics, such as allowing persistent experimentation and limited memory. Model predictions are derived from computer simulation of the reinforcement process. The intermediate run of the average simulation path is compared to the actual path observed in the experiment.

Roth and Erev emphasize that the model possesses some of the robust properties of learning long noted in the psychology literature. These are the Law of Effect, which states that choices that have led to good outcomes in the past are more likely to be repeated in the future; and the Power Law of Practice, which states that learning curves tend to be steep at the beginning, and then flatten out.

Consistent with the Law of Effect, the basic mechanics of the model imply that the higher the payoff received from playing a pure strategy, the higher the probability that the pure strategy will be played in the next round. As a consequence, for the ultimatum game, second movers learn to accept smaller offers more slowly than larger offers. This drives first movers away from very small offers, towards the more acceptable larger ones. As the number of small offers dwindles, play stabilizes away from perfect equilibrium. The key to the result, in a nutshell, is that second movers learn not to reject very small offers more slowly than first movers learn not to make them.

Roth and Erev also consider the public goods game known as best shot. In this two-person game, the first mover chooses a contribution of tokens. After viewing the first mover's choice, the second mover makes a contribution. The maximum of the two determines the total cash to be divided, the larger the maximum, the larger the pie. The largest payoff, however, goes to the player who contributes the least. Perfect equilibrium has the second mover doing all the contributing, leading to an allocation that looks strikingly like that for the ultimatum game, with the second mover receiving much less than the first mover. Along the equilibrium path, the best shot second mover has the opportunity to end the game with both players receiving nothing, just like the ultimatum game second mover.

Despite the similarities, best shot experiments produce very different results than ultimatum experiments. Studies by Harrison and Hirshleifer (1989) and Prasnikar and Roth (1992) find that after a few iterations, best shot play approaches 100% perfect equilibrium. It is an outcome that the Roth-Erev model anticipates. The key is that best shot and ultimatum games are quite different off-the-equilibrium path. Best shot allows the second mover to exploit a generous (off-the-equilibrium path) first mover contribution in a way that leaves the first mover with little. No such option is available to the ultimatum second mover. The reinforcement model predicts that best shot second movers quickly learn to exploit any first mover generosity. So unlike ultimatum, first movers in best shot move towards the equilibrium strategy. As a result, best shot second movers receive more persistent reinforcement to take the small amount equilibrium allots them than do ultimatum second movers. Hence reinforcement learning pushes both best shot players towards the subgame perfect equilibrium. Roth and Erev show that their model also predicts the subgame perfect equilibrium behavior observed in a simple market game.

Duffy and Feltovich (forthcoming) provide further evidence for reinforcement learning for ultimatum and best shot games. Miller and Andreoni (1991) demonstrate that replicator dynamics can explain some of the stylized facts from public goods experiments.

The Roth and Erev model does not attempt to predict initial playing propensities. These are taken as given.¹⁰ How we interpret these initial propensities is nevertheless important when we go to compare reinforcement learning to motivation and choice models. The point is perhaps most clearly made in the context of second movers in the ultimatum game. How do we interpret second movers' initial propensity to turn down money?

It seems to me there are two possible answers. One is that at least some responders begin the game confused about the implications of playing reject. Experience clarifies things. But this interpretation seems to run contrary to the strong negative correlation observed between rejecting and the size of the offer (e.g., Bolton and Zwick, 1995). If it were just confusion, we would expect no correlation. A second explanation, suggested by both motivation and choice models, does not have this problem; it is of course that second movers reject when they find an offer unacceptably unfair. Taking this as our interpretation, the crucial difference between learning on the one hand, and motivation and choice on the other, is that learning assumes that the concern for fairness will dissipate with enough reinforcement of the right sort.

Learning is without a doubt observed in some of these games.¹¹ Recall that McKelvey and Palfrey (1992) found evidence for learning in the centipede game. Winter and Zamir (1997) study learning behavior by matching real ultimatum game players with computer players. They find that first movers search for the optimum offer in much the way we would expect from reinforcement learning. But the same experiment casts doubt on the idea that second movers learn; specifically, Winter and Zamir find that second mover rejection habits change little with repeated play even when second movers are matched with computer programs that persistently make low offers.

Hence it does not appear that learning models are substitutes for motivation or choice models. Nor vice versa: Motivation and choice models are static, and so cannot account for learning. An experiment by Abbink et al. (1996) attempts to pinpoint precisely where learning and motivation explanations are necessary for the ultimatum game. They find that motivation explanations are necessary to explain both initial rejection propensities and the persistence of rejection behavior. On the other hand, they find evidence that first movers learn in a manner consistent with what we might expect from reinforcement learning.

7. Evolutionary models: Explaining why

So far I have described theories that attempt to explain *how* behavior is bounded away from full rationality. None of these theories attempts to explain *why* rationality is limited in the particular way asserted. While the *why* question is not the focus of this paper, I think the *why*-models provide an important foundation for the *how*-models, and so I want to make at least brief mention of them here.

An answer to the *why*-question is important for at least three reasons. First, an empirically validated *why*-answer would increase our confidence in the associated *how*-approach. Second, the *why*-answer may give us a clue as to what directions, and to what limits, the *how*-approach can be extended. Third, a *why*-answer would satisfy the curiosity of the reductionists among us, who prefer answers expressed in generally recognized primitives. The models in this section argue that the appropriate primitives for fairness and reciprocity are biological in nature.

7.1. *Evolution and social learning*

The social learning approach posits that subjects come into the lab and apply what they know from “the game of life,” lessons that may or may not be entirely appropriate to the particular lab game. Gale et al. (1995) develop this idea formally. They apply an evolutionary dynamic that is subject to constant perturbation or ‘noise’ to the ultimatum game. The modelers “expect the likelihood of a learning error to depend on how much it currently matters in payoff terms what strategy is played in the game,” meaning that second movers who receive small offers tend to be noisier than the first movers. Simulations of the dynamic produce outcomes that settle down away from subgame perfect equilibrium, and are quantitatively similar to those observed in the lab. The model suggests that second movers fail to accept small sums of money because the biological penalty for doing so is too slight for the evolutionary process to have extinguished it.

Eshel et al. (1998) do an evolutionary analysis of a public goods game in which people are either altruists or self-interested egoists. In this model, people learn by imitation, and they interact locally. Eshel et al. show that, under these circumstances, altruists can survive if they are grouped together.

7.2. *Indirect evolutionary approach*

We can think of rejecting money as an attempt to impose a penalty on someone you feel has treated you unfairly—that is, as an act of negative reciprocity. Cooperating in a prisoner’s dilemma in exchange for the other player cooperating is then a positive reciprocal act. Indirect evolutionary models attempt to explain why the evolutionary process might favor a preference for reciprocal acts, both positive and negative. Güth and Yaari (1992) introduce the indirect evolutionary approach, in which it is assumed that evolutionary forces shape utility functions, but given these utility functions, people act rationally. So in the indirect evolutionary approach, people are not fully rational, but nor are they strictly stupid.

Güth (1995) presents a detailed indirect evolutionary model of negative reciprocity. Huck and Oechssler (forthcoming) use an indirect evolutionary argument to derive the utility functions used in Bolton’s (1991) bargaining model. Kockesen et al. (1997) look at the evolutionary performance of interdependent preferences in common pool resource games and public goods games.

7.3. *Evolutionary psychology*

The previous models have been posed in terms of evolutionary game theory. There is also a developed literature in the field of evolutionary psychology. The hypothesis from that literature that is relevant here is that humans have evolved mental algorithms for identifying and punishing cheaters who behave non-cooperatively in social exchange. Hoffman et al. (1998) apply some of the algorithms from this literature to various bargaining and dilemma games.

All these models face a common challenge: testability. One way to get some sense of whether or not they are on the right track is to look for animal equivalents of the behavior

we see in the economics lab. For example, de Waal (1996) argues that primitive versions of fairness and morality are found in animal societies. He describes what is essentially a dictator game experiment performed on capuchin monkeys (p. 148). Two monkeys were placed in adjacent cages with a mesh partition between them. One monkey was given food, and sharing was often observed. While there are several possible explanations for precisely why the monkeys share, the fact that they share at all lends some support to the hypothesis that the behavior in the human dictator game has a biological foundation.

8. Summary: Robustness

Many factors influence human behavior. It is probably impossible to capture all of them in a single model. Even if it were possible, the resulting model would lack parsimony, and would almost certainly be too opaque to be informative. Practically speaking, a model must focus on a few select factors. Each of the models discussed here focuses on a somewhat different set, and each can point to some evidence in its favor. But so long as a model must focus on a few select factors, it is bound to fail sometimes, if only under the circumstances where the omitted factors are the important ones. For this reason, I would say that empirically testing a model is less about determining whether the model is true or false, and more about assessing the model's robustness.

Robust in two senses: For one, robust in the sense of being accurate over a broad domain, or if you will, over a wide class of games. Second, robust in the sense of having economic, as well as statistical, significance; that is, the factors the model accounts for should have a meaningful quantitative influence. The two types of robustness do not necessarily come as a package. Framing—how the game is explained to subjects—appears to have a broad domain of influence in that framing effects have been reported for many experiments. In a few cases, the framing effect is large; in others, the effect is statistically significant but nonetheless small (see end note 4 for some examples). So while framing is clearly broadly influential, I think its economic significance is, at this point, more difficult to judge.

Developing and testing on a wider sphere of games is a rigorous way of demonstrating just how broad a model's grasp is. Hence the importance of finding ways to bridge gaps—such as the one between bargaining and dilemma games. Efforts to establish this sort of robustness for some of the models discussed here are now under way.

The issue of economic significance is also being played out. For example, the question of whether it is necessary to account for intentions, or whether it is sufficient to confine attention to distribution is essentially an economic significance issue. Intentions matter, but do they matter enough to justify the additional theoretical machinery necessary to capturing them? Of course the answer depends, at least in part, on the preponderance of data. But I purposely phrase the question in a manner that suggests a trade-off, because I suspect the answer will also depend on the ability of theorists to pose parsimonious models of the influence of distribution, on the one hand, and of intentions on the other. The simpler the theoretical framework, the lower the threshold on the benefits necessary to justify the model's carrying costs.

The finitely repeated prisoner's dilemma, the experiment we began with, presents a related issue. Recall that Andreoni and Miller found evidence for both reputation building

and altruism. Selten and Stoecker, on the other hand, explain their experiment in terms of learning. McKelvey and Palfrey study a closely related dilemma, the centipede game, and find evidence that all three factors are important, although their own model formally concerns just two, reputation building and altruism. There would seem then, a fair amount of evidence that all three—reputation building, altruism, and learning—play a role in these types of games. It might be nice to have a model that embraces all three, if only to shed some light on how they interact.

But, more generally, are we looking for one grand theory capable of explaining all the significant phenomena in all the games we are interested in? I would say probably not, if only because such a theory sounds awfully cumbersome. It would probably be better to have a set of building block models that can be mixed and matched as needed. Explaining the ultimatum game might require different building blocks than explaining the finitely repeated prisoner's dilemma. Public goods games and the centipede game might use the same blocks, but in somewhat different configurations.

Let me suggest three building blocks: Motivation, a basic theory of what motivates people under different circumstances. Learning, an explanation of how people adjust to experience and new information. And finally, strategic reasoning. All of the models here depart from classical, full rationality game theory in some way, but most nevertheless borrow substantially from it. To me there is little doubt that the analytical ins-and-outs of a game strongly influence the behavior we see, often from the very first round of play (which is why I think strategic reasoning is something different than learning). The challenge is to construct a model of strategic reasoning from principles that are consistent with the capability and limitations of the human mind.

Hence, the finished product may be a set of separate components, each with a distinct function. Components would then be custom matched and configured depending on the task at hand. It is perhaps a more elaborate vision of theory than the one with which experimental economists began their study of bargaining and dilemma games. But the data seems to demand a more elaborate theory. Parsimony if at all possible, but no sacrificing the data.

Acknowledgment

This paper elaborates on a presentation I made at The Bonn Workshop on Theories of Bounded Rationality. I thank Elena Katok, Axel Ockenfels and two anonymous referees for many helpful comments.

Notes

1. In the one-shot prisoner's dilemma, two players each simultaneously choose to either 'cooperate' or 'defect'. Using the payoffs in the Andreoni and Miller experiment, if both cooperate, then each receives a payoff of 7. If both defect, each receives 4, and if one cooperates and the other defects, then the defector receives 12 and the cooperator receives 0. For the one-shot prisoner's dilemma, defecting is a dominant strategy; that is, regardless of what a player's partner does, the player receives a higher payoff by defecting.
2. It might be argued that the experiment indicates there is something wrong with the subgame perfect equilibrium refinement, rather than the more basic concept of Nash equilibrium. However, McKelvey and Palfrey's (1992)

experiment, which I discuss in Section 6, exhibits a similar pattern of “cooperation” in a game where all standard solution concepts predict we should see none.

3. One experiment where dictators did play two sequential dictator games is Cason and Mui (1998). The two treatments were distinguished by the information given to dictators about other dictators prior to the second play of the game. There was a decrease in second round giving in one treatment, but not in the other treatment.
4. Hoffman et al. (1994) report two treatments that are subject-experimenter anonymous. Dictator giving is 9.2 and 10.5% of the sum, respectively. Bolton et al. (1998) report a treatment in which giving is 13.5%.
5. ‘Framing effect’ refers to an influence on subject behavior due to the way the experimenter phrases the instructions. Bolton et al. (1998) compare the results of three independent dictator experiments. They find that a treatment run by Hoffman et al. (1994) in a buyer-seller frame is not statistically different from the Forsythe et al. result of figure 2. A second Hoffman et al. treatment run in a contest frame is weakly significantly different than the Forsythe et al. treatment. Hoffman et al.’s double blind (subject-experimenter anonymous) treatments are strongly statistically different than Forsythe et al. Pruitt (1967) reports framing effects in the context of the prisoner’s dilemma. Andreoni (1995b) demonstrates framing effects in a public goods game.
6. Camerer (1997) uses a similar taxonomy to categorize work in behavioral game theory.
7. Other’s utility is sometimes used instead. But from an empiricist’s point of view, other’s consumption is the preferred variable since it alone is observable.
8. Levine (1998) analyzes an extensive form model that has a somewhat similar flavor.
9. See the discussion (p. 113) and accompanying references in Wilson (1997).
10. In their paper, Roth and Erev (1995) first analyze the model using initial conditions drawn randomly from a uniform distribution, and then compare the model to experimental data using initial conditions fitted from the first rounds of play.
11. Camerer and Ho (forthcoming) develop a learning model in which belief-based learning, where players act on beliefs formed about what others will do in the future based on past observation, and reinforcement learning are special cases. They show that the model fits significantly better than others for a variety of games.

References

- Abbinck, Klaus, Bolton, Gary E., Sadrieh, Abdolkarim, and Tang, Fang-Fang. (1996). “Adaptive Learning versus Punishment in Ultimatum Bargaining.” Working Paper, University of Bonn.
- Anderson, Simon P., Goeree, Jacob K., and Holt, Charles A. (forthcoming). “A Theoretical Analysis of Altruism and Decision Error.” *Journal of Public Economics*.
- Andreoni, James. (1989). “Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence.” *Journal of Political Economy*. 97, 1447–1458.
- Andreoni, James. (1993). “An Experimental Test of the Public-Goods Crowding-Out Hypothesis.” *American Economic Review*. 83, 1317–1327.
- Andreoni, James. (1995a). “Cooperation in Public Goods Experiments: Kindness or Confusion?” *American Economic Review*. 85, 891–904.
- Andreoni, James. (1995b). “Warm-Glow versus Cold-Pickle: The Effects of Positive and Negative Framing on Cooperation in Experiments.” *Quarterly Journal of Economics*. 110, 1–22.
- Andreoni, James and Miller, John H. (1993). “Rational Cooperation in the Finitely Repeated Prisoner’s Dilemma: Experimental Evidence.” *Economic Journal*. 103, 570–585.
- Aumann, Robert J. and Maschler, Michael. (1964). “The Bargaining Set for Cooperative Games.” In M. Dresher, L.S. Shapley, and A.W. Tucker (eds.), *Advances in Game Theory*. Princeton: Princeton University Press, pp. 443–476.
- Bacharach, Michael and Bernasconi, Michele. (1997). “The Variable Frame Theory of Focal Points: An Experimental Study.” *Games and Economic Behavior*. 19, 1–45.
- Becker, Gary S. (1974). “A Theory of Social Interactions.” *Journal of Political Economy*. 82, 1063–1093.
- Berg, Joyce, Dickhaut, John, and McCabe, Kevin. (1995). “Trust, Reciprocity, and Social Norms.” *Games and Economic Behavior*. 10, 122–142.
- Bernheim, B. Douglas. (1986). “On the Voluntary and Involuntary Provision of Public Goods.” *American Economic Review*. 76, 789–793.

- Binmore, Kenneth, Shaked, Avner, and Sutton, John. (1985). "Testing Noncooperative Bargaining Theory: A Preliminary Study." *American Economic Review*. 75, 1178–1180.
- Blount, Sally. (1995). "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes*. 63, 131–144.
- Bolton, Gary E. (1991). "A Comparative Model of Bargaining: Theory and Evidence." *American Economic Review*. 81, 1096–1136.
- Bolton, Gary E., Brandts, Jordi, and Ockenfels, Axel. (1998). "Measuring Motivation in the Reciprocal Responses Observed in a Dilemma Game." *Experimental Economics*, this issue.
- Bolton, Gary E. and Katok, Elena. (1998). "An Experimental Test of the Crowding Out Hypothesis: The Nature of Beneficent Behavior." *Journal of Economic Behavior and Organization*. 37, 315–331.
- Bolton, Gary E., Katok, Elena, and Zwick, Rami. (1998). "Dictator Game Giving: Rules of Fairness versus Acts of Kindness." *International Journal of Game Theory*. 27, 269–299.
- Bolton, Gary E. and Ockenfels, Axel. (1997). "ERC: A Theory of Equity, Reciprocity and Competition." Working Paper, Penn State University.
- Bolton, Gary E. and Ockenfels, Axel. (1998). "Strategy and Equity: An ERC-Analysis of the Güth-van Damme game." *Journal of Mathematical Psychology*. 42, 215–226.
- Bolton, Gary E. and Zwick, Rami. (1995). "Anonymity versus Punishment in Ultimatum Bargaining." *Games and Economic Behavior*. 10, 95–121.
- Brandts, Jordi and Schram, Arthur. (1996). "Cooperative Gains or Noise in Public Goods Experiments." Working Paper, Tinbergen Institute.
- Budescu, David V., Rapoport, Amnon, and Suleiman, Ramzi. (1995). "Common Pool Resource Dilemmas under Uncertainty: Qualitative Tests of Equilibrium Solutions." *Games and Economic Behavior*. 10, 171–201.
- Camerer, Colin F. (1997). "Progress in Behavioral Game Theory." *Journal of Economic Perspectives*. 11, 167–188.
- Camerer, Colin F. and Ho, Teck-Hua. (forthcoming). "Experience-weighted Attraction Learning in Normal-form Games." Working Paper, California Institute of Technology, *Econometrica*.
- Cason, Timothy N. and Mui, Vai-Lam. (1998). "Social Influence in the Sequential Dictator Game." *Journal of Mathematical Psychology*. 42, 248–265.
- Charness, Gary. (1996). "Attribution and Reciprocity in a Simulated Labor Market: An Experimental Investigation." Working Paper, UC Berkeley.
- de Waal, Frans B.M. (1996). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, MA: Harvard University Press.
- Duffy, John and Feltovich, Nick. (forthcoming). "Does Observation of Others Affect Learning in Strategic Environments? An Experimental Study." *International Journal of Game Theory*.
- Eshel, Ilan, Samuelson, Larry, and Shaked, Avner. (1998). "Altruists, Egoists, and Hooligans in a Local Interaction Model." *American Economic Review*. 88, 157–179.
- Fehr, Ernst, Kirchsteiger, Georg, and Riedl, Arno. (1993). "Does Fairness Prevent Market Clearing? An Experimental Investigation." *Quarterly Journal of Economics*. 108, 437–459.
- Fehr, Ernst and Schmidt, Klaus. (1997). "A Theory of Fairness, Competition and Cooperation." Working Paper, University of Zurich.
- Forsythe, Robert, Horowitz, Joel L., Savin, N.E., and Sefton, Martin. (1994). "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior*. 6, 347–369.
- Fudenberg, Drew and Levine, David K. (1997). "Measuring Subject's Losses in Experimental Games." *Quarterly Journal of Economics*. 112, 508–536.
- Gale, John, Binmore, Kenneth G., and Samuelson, Larry. (1995). "Learning to be Imperfect: The Ultimatum Game." *Games and Economic Behavior*. 8, 56–90.
- Geanakoplos, John, Pearce, David, and Stacchetti, Ennio. (1989). "Psychological Games and Sequential Rationality." *Games and Economic Behavior*. 1, 60–79.
- Güth, Werner. (1995). "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives." *International Journal of Game Theory*. 24, 323–344.
- Güth, Werner and van Damme, Eric. (1998). "Information, Strategic Behavior and Fairness in Ultimatum Bargaining: An Experimental Study." *Journal of Mathematical Psychology*. 42, 227–247.
- Güth, Werner, Schmittberger, Rolf, and Schwarze, Bernd. (1982). "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization*. 3, 367–388.

- Güth, Werner and Yaari, Menahem. (1992). "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game." In U. Witt (ed.), *Explaining Process and Change—Approaches to Evolutionary Economics*. Ann Arbor: The University of Michigan Press, pp. 23–34.
- Harrison, Glenn W. and Hirshleifer, J. (1989). "An Experimental Evaluation of Weakest Link/Best Shot Models of Public Goods." *Journal of Political Economy*. 97, 201–225.
- Hochman, Harold M. and Rodgers, James D. (1969). "Pareto Optimal Redistribution." *American Economic Review*. 59, 542–557.
- Hoffman, Elizabeth, McCabe, Kevin, Shachat, Keith, and Smith, Vernon L. (1994). "Preferences, Property Rights and Anonymity in Bargaining Games." *Games and Economic Behavior*. 7, 346–380.
- Hoffman, Elizabeth, McCabe, Kevin, and Smith, Vernon. (1998). "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology." *Economic Inquiry*. 36, 335–352.
- Huck, Steffen and Oechssler, Jörg. (forthcoming). "The Indirect Evolutionary Approach to Explaining Fair Allocations." *Games and Economic Behavior*.
- Isaac, R. Mark and Walker, James M. (1988). "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism." *Quarterly Journal of Economics*. 103, 179–199.
- Isaac, R. Mark, Walker, James M., and Williams, Arlington W. (1994). "Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups." *Journal of Public Economics*. 54, 1–36.
- Kagel, John, Kim, Chung, and Moser, Donald. (1996). "Fairness in Ultimatum Games with Asymmetric Information and Asymmetric Payoffs." *Games and Economic Behavior*. 13, 100–110.
- Kahan, James P. and Rapoport, Amnon. (1974). "Test of the Bargaining Set and Kernel Models in Three-Person Games." In Am. Rapoport (ed.), *Game Theory as a Theory of Conflict*. Dordrecht. Holland: D. Reidl, pp. 119–160.
- Kahneman, Daniel, Knetsch, Jack L., and Thaler, Richard H. (1986). "Fairness and the Assumptions of Economics." *Journal of Business*. 59, S285–S300.
- Kockesen, Levent, Ok, Efe A., and Sethi, Rajiv. (1997). "Interdependent Preference Formation." Working Paper, Barnard College.
- Kreps, David M., Milgrom, Paul, Roberts, John, and Wilson, Robert. (1982). "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma." *Journal of Economic Theory*. 27, 245–252.
- Kuon, Bettina and Uhlich, Gerald R. (1993). "The Negotiation Agreement Area: An Experimental Analysis of Two-Person Characteristic Function Games." *Group Decision and Negotiation*. 2, 323–345.
- Ledyard, John. (1995). "Public Goods: A Survey of Experimental Research." In John H. Kagel and Alvin E. Roth (eds.), *The Handbook of Experimental Economics*. Princeton: Princeton University Press, pp. 111–194.
- Levine, David K. (1998). "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*. 1, 593–622.
- McKelvey, Richard D. and Palfrey, Thomas R. (1992). "An Experimental Study of the Centipede Game." *Econometrica*. 60, 803–836.
- McKelvey, Richard D. and Palfrey, Thomas R. (1995). "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior*. 10, 6–38.
- Miller, John H. and Andreoni, James (1991). "Can Evolutionary Dynamics Explain Free Riding in Experiments?" *Economics Letters*. 36, 9–15.
- Palfrey, Thomas R. and Prisbrey, Jeffrey E. (1997). "Anomalous Behavior in Public Goods Experiments: How Much and Why?" *American Economic Review*. 87, 829–846.
- Prasnikar, Vesna and Roth, Alvin E. (1992). "Considerations of Fairness and Strategy: Experimental Data from Sequential Games." *Quarterly Journal of Economics*. 107, 865–888.
- Pruitt, D.G. (1967). "Reward Structure and Cooperation: The Decomposed Prisoner's Dilemma Game." *Journal of Personality and Social Psychology*. 7, 21–27.
- Rabin, Matthew. (1993). "Incorporating Fairness into Game Theory and Economics." *American Economic Review*. 83, 1281–1302.
- Roth, Alvin E. (1985). "Towards a Focal Point Theory of Bargaining." In A.E. Roth (ed.), *Game-Theoretic Models of Bargaining*. Cambridge: Cambridge University Press, pp. 259–268.
- Roth, Alvin E. (1995). "Bargaining Experiments." In J. Kagel and A.E. Roth (eds.), *The Handbook of Experimental Economics*. Princeton: Princeton University Press, pp. 253–348.

- Roth, Alvin E. and Erev, Ido. (1995). "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term." *Games and Economic Behavior*. 8, 164–212.
- Roth, Alvin E., Malouf, Michael W.K., and Murnighan, J. Keith. (1981). "Sociological versus Strategic Factors in Bargaining." *Journal of Economic Behavior and Organization*. 2, 153–177.
- Schelling, Thomas C. (1960). *The Strategy of Conflict*. Cambridge: Harvard University Press.
- Selten, Reinhard. (1987). "Equity and Coalition Bargaining in Experimental Three-person Games." In A.E. Roth (ed.), *Laboratory Experimentation in Economics: Six Points of View*. Cambridge: Cambridge University Press, pp. 42–98.
- Selten, Reinhard and Krischker, W. (1982). "Comparison of Two Theories for Characteristic Function Experiments." In R. Tietz (ed.), *Aspiration Levels in Bargaining and Economic Decision Making*. Springer Lecture Notes in Economic and Mathematical Systems No. 213, Berlin: Springer-Verlag, pp. 259–264.
- Selten, Reinhard and Ockenfels, Axel. (1998). "An Experimental Solidarity Game." *Journal of Economic Behavior and Organization*. 34, 517–539.
- Selten, Reinhard and Stoecker, Rolf. (1986). "End Behaviour in Sequences of Finite Prisoner's Dilemma Supergames: A Learning Theory Approach." *Journal of Economic Behavior and Organization*. 7, 47–70.
- Sugden, Robert. (1984). "Reciprocity: The Supply of Public Goods through Voluntary Contributions." *Economic Journal*. 94, 772–787.
- Van Huyck, John B., Battalio, Raymond C., and Walters, Mary F. (1995). "Commitment versus Discretion in the Peasant-Dictator Game." *Games and Economic Behavior*. 10, 143–170.
- Wilson, Edward O. (1997). *Consilience*. New York: Alfred A. Knopf, Inc., opening quote from p. 96.
- Winter, Eyal and Zamir, Shmuel. (1997). "An Experiment with Ultimatum Bargaining in a Changing Environment." Working Paper, Washington University.