Analyse & Kritik 27/2005 (© Lucius & Lucius, Stuttgart) 48–72

Herbert Gintis

Behavioral Game Theory and Contemporary Economic Theory

Abstract: It is widely believed that experimental results of behavioral game theory undermine standard economic and game theory. This paper suggests that experimental results present serious theoretical modeling challenges, but do not undermine two pillars of contemporary economic theory: the rational actor model, which holds that individual choice can be modeled as maximization of an objective function subject to informational and material constraints, and the incentive compatibility requirement, which holds that macroeconomic quantities must be derived from the interaction and aggregation of individual choices. However, we must abandon the notion that rationality implies self-regarding behavior and the assumption that contracts are costlessly enforced by third parties.

1. Introduction

The articles that serve as the focus of this Symposium on altruism are among the best of a new genre. The genre is *behavioral game theory*, which may be loosely defined as the application of game theory to the design of laboratory experiments. Behavioral game theory aims to determine empirically how individuals make choices under conditions of uncertainty and strategic interaction. It is widely believed that experimental results of behavioral game theory undermine standard economic and game theory. This paper suggests that experimental results present serious theoretical modeling challenges, but do not undermine two pillars of contemporary economic theory: the *rational actor model*, which holds that individual choice can be modeled as maximization of an objective function subject to informational and material constraints, and the *incentive compatibility* requirement, which holds that macroeconomic quantities must be derived from the interaction and aggregation of individual choices. However, we must abandon the notion that rationality implies self-regarding behavior and the assumption that contracts are costlessly enforced by third parties.

Behavioral game theory can be roughly divided into five interdependent and partially overlapping stages. The first consists of the Ellsberg, Allais and related paradoxes, which suggest that probabilities enter in a nonlinear manner into the determination of expected utility (Allais 1953; Ellsberg 1961). Allais was awarded the Nobel prize in 1988. Segal (1987), Machina (1987), and others have shown that this behavior can be analytically modeled as expected utility with nonlinear weights.

The second wave of behavioral game theoretic results is exemplified by the research of Vernon Smith and his coworkers. Smith was awarded the Nobel prize in Economics in 2002. Smith began running laboratory experiments of market exchange in 1956 at Purdue and Stanford Universities. His pioneering results strongly supported the model of the rational, self-interested actor and of price-equilibrated market exchange.

The third stage consists of the contributions of Amos Tversky, Daniel Kahneman and their coworkers to behavioral decision theory beginning in the early 1970's, culminating in Kahneman's being awarded the Nobel prize in Economics in 2002, the same year as Vernon Smith. Kahneman's and Tversky's work is a sustained empirical critique of traditional decision theory. This impressive body of research has led to several substantive models of decision-making outside the standard model developed by Von Neumann and Morgenstern (1944), Savage (1954), di Finetti (1974) and others, including prospect theory (Kahneman/Tversky 1979), hyperbolic discounting (Ainslie/Haslam 1992; Ahlbrecht/ Weber 1995; Laibson 1997), and regret theory (Sugden 1993). While integrating these alternatives into the larger economic frameworks of market exchange and economic regulation presents considerable analytical modeling challenges, they not incompatible with maximization subject to constraints.

The fourth stage includes the ultimatum game research of Güth et al. (1982), the bargaining experiments of Roth and his coworkers (Roth et al. 1991; Roth 1995), the trust game research of Berg et al. (1995), and the common pool resource and public goods studies of Elinor Ostrom, Toshio Yamagishi and their coworkers (Yamagishi 1986; Hayashi et al. 1999; Watabe et al. 1996). These represent the first systematic investigation of decision-making under conditions of strategic interaction. A characteristic of this fruitful period of research is that experimenters generally consider the non-self-interested behavior of agents as anomalous and based on irrational behavior and faulty reasoning on the part of subjects.

The fifth, and most recent, stage in behavioral game theory research consists of the various experimental scenarios investigated by Ernst Fehr and his coworkers featured here, along with related contributions by his coworkers, as well as Levine (1998), Bolton and Ockenfels (2000), Charness and Rabin (2002) and others. These contributions sharpen and extend the finding of the fourth stage, but present a challenge of quite a different order. Rather than treating anomalous behavior as faulty reasoning or behavior, they build analytical models premised upon the rational decision theory, but with agents who systematically exhibit *other-regarding preferences*; i.e., they care about not only their own payoffs in a strategic interaction, but those of the other players as and the process of play well.

In this introduction to the symposium, I will address some general issues to which this 'fifth wave' of research has given rise. I will argue the following points:

a. **Expanding the Rational Actor Model.** This fifth wave research supports the 'thin' concept of rationality on which contemporary decision theory, game theory, and microeconomic theory are based. This conception assumes only that preferences are consistent over the appropriate

choice space. Other-regarding preferences do, however, expand the content of the preference function beyond the traditional exclusive reliance on personal gain through consumption, leisure and asset portfolio enhancement. Moreover, the proper choice space must be empirically determined. For instance, according to prospect theory (Kahneman/Tversky 1979), the choice space privileges the agent's current position, and in hyperbolic discounting the choice space privileges the time at which choice is exercised (Ahlbrecht/Weber 1995).

b. **Other-Regarding Preferences.** Several categories of other-regarding preferences need be added to the standard model to capture human behavior. These include strong reciprocity, inequality aversion, and 'insider' bias. We define these as follows. A *social dilemma* is a game with two pure strategies, 'cooperate' and 'defect' in which all other players gain when a player cooperates, but a self-regarding player will always defect, giving no benefit to the group, whatever the other players do. Strong reciprocity is a predisposition to cooperate in a social dilemma, and to punish non-cooperators when possible, at a personal cost that cannot be recouped in later stages of the game. Inequality aversion is the predisposition to reduce the inequality in outcomes between oneself and other group members, even at personal cost. Insider bias in a game is the predisposition to identify other players who are 'like oneself' according to some payoff-irrelevant ascriptive marker (such as ethnicity or nationality) and behave altruistically on behalf of these 'insiders'. These categories are probably universal, but their content is culturally variable. They are supported by such psychological traits as the capacity to internalize social values, and the tendency to display such social emotions as empathy, shame, pride, and remorse.

c. **Complete Contracting.** A *complete contract* among a group of agents is an agreement specifying the rights and obligations of each party under all possible future states of affairs, costlessly written and enforced by third parties (e.g., the judiciary). In anonymous competitive market settings with complete contracting, individuals behave like the self-regarding actor of traditional economic theory.

d. **Incomplete Contracting.** A *one-sided* incomplete contract is one in which one party to an exchange delivers a contractually enforceable quantity (e.g., money) in return for an unenforceable promise of delivery of services (e.g., work). Under conditions of competitive market exchange with one-sided incomplete contracting, other-regarding preferences (gift exchange, conditional cooperation and punishment) emerge. Such situations often attain a high level of allocational efficiency compared to the situation with self-regarding agents. These situations are characterized by non-clearing markets in which the agent on the short side of a con-

tractual relationship, usually the party who is offering money, has power in some meaningful, quasi-political sense (e.g., employers, lenders, consumers) while agents on the long side enjoy rents (employees, borrowers, firms).

Section 2 explains why other-regarding preferences enrich rather than undermine rational choice theory. The reason is that rational choice theory requires only that preferences be *consistent*, and is in principle agnostic to the *content* of preferences. This should be completely obvious to economists, but the epithet "irrational" is so frequently applied in a manner inconsistent with its proper use in economic theory that formally addressing this issue appears to be in order. The upshot is that we can continue to affirm the principle that agents can be successfully modeled as maximizing a preference function subject to informational and material constraints.

Section 3 explores the implications of experimental economics for game theory. Since game theory provides the methodological foundations for experimental design and analysis in experimental economics, if the latter's empirical findings undermined game theory, they would thereby undermine their own validity—a situation demanding a serious, radical reconstruction of the general theory of strategic interaction. In fact, however, since the rational choice theory remains intact, we can assume agents choose best responses in strategic interactions, and hence game theory is not undermined. Some experimental research, however, does suggest that game-theoretic predictions involving more than a few levels of backward induction on the part of agents generally predict very poorly, suggesting that agents do not choose best responses, and hence game theory itself is threatened (McKelvey/Palfrey 1992; Camerer 2003). An important branch of game theory, known as *interactive decision theory*, often overlooked in methodological discussions of the implications of empirical research, indicates however that backward induction can be identified with choosing best responses *only under specialized conditions*, or only making questionable assumptions concerning the nature of logical and statistical inference (Fagin et al. 1995; Halpern 2001; Aumann 1995; Aumann/Brandenburger 1995). It follows that the experimental findings on backward induction do not threaten game theory, although they counsel against the indiscriminate use of backward induction arguments in parts of the game tree that cannot be reached by rational agents.[1]

Additional support for traditional economic theory comes from the fact that when all aspects of market exchange are covered by complete contracts, agents behave as self-interested income maximizers, as suggested in traditional economic theory. Many experiments carried out by Vernon Smith and his coworkers support this generalization, and in Section 4, we present recent, relatively elaborate, studies that come to the same conclusion.

Many of the characteristics of modern market economies are the result of *incomplete contracting*. Gintis (1976) suggested that the major outlines of the employer-employee relationship (long-term contracts with supra-market-clearing

[1] Many weaknesses of classical game theory are overcome using evolutionary game theory. I direct the reader to Gintis 2000.

wages, job ladders, and the use of promotion and dismissal as motivating devices) are due to the fact that in return for a wage, the worker cannot credibly guarantee any particular level of effort or care in the labor-time provided the employer. Akerlof (1982) suggested that under such conditions the employer-employee relationship could be a 'gift exchange' situation, in which workers voluntarily supply a high level of effort when they believe that their employer is offering a fair wages and good working conditions. Bowles and Gintis (1993) introduced the notion of *short-side power* in the following terms: "The short side of a market is the side for which the quantity of desired transactions is the least. Short-side agents include employers in labor markets with equilibrium unemployment, ... and lenders in capital markets with equilibrium credit rationing." We asserted the following principle: "competitive equilibrium ... allocates power to agents on the short side of non-clearing markets". In particular, there tend to be both job rationing and credit rationing, in the sense that there are always more applicants for a job than job openings, and this excess supply does not lead to a bidding down of wages. Similarly, there are more applicants for loans than there are loanable funds, and this excess demand leads to strong collateral requirements rather than the bidding up of the interest rate. Gintis (1989) applied a similar argument to the relationship between consumers and firms that supply goods where contracts do not ensure the delivery of high quality products. In this case, the supplying firm is on the long side of the market (sellers are quantity constrained), and price is higher than marginal cost, accounting for the fact that many firms in a market economy see their task as 'selling their product', rather than maximizing profits with a given demand function.

Section 5 describes the achievement of Ernst Fehr, Simon Gächter and Georg Kirchsteiger (1997) in showing that Akerlof's gift exchange mechanism is strongly operative when the labor contract is incomplete. In a more elaborate setting, Martin Brown, Armin Falk, and Ernst Fehr (2004) show that both gift exchange and threat of dismissal are operative in incomplete contract setting. This experimental setting, described in Section 6, is especially interesting because it illustrates the coexistence of self- and other-regarding incentives in a single game. While doubtless at times self-regarding incentives 'crowd out' other-regarding motives (Frey 1997a;b), at least in the labor market the two probably coexist. While there have been several attempts at interpreting these results in such manner as to preserve the assumption of self-regarding behavior, I am convinced that they fail. I develop this argument in Section 7.

2. Rational Choice Theory

Rational choice theory models behavior as agents maximizing a preference function subject to informational and material constraints. The term "rational" is a misnomer, since the term appears to imply something about the ability of the agent to give reasons for actions, to act objectively, unmoved by capricious emotionality, and even to act self-interestedly. Yet, it has long been recognized that this connotational overlay is superfluous and misleading. Nothing has brought

this fact home more clearly than the great success of the rational actor model in explaining animal behavior, despite the fact that no one believes that fruit flies and spiders do much in the way of cogitating (Maynard Smith 1982; Alcock 1993). Rational choice theory is the starting point for much of economic analysis, behavioral game theory, and is increasingly gaining credence with neuroscientists (Shizgal 1999; Glimcher 2003).

Formally, the assertion that consistent preferences are sufficient to model the individual as maximizing a preference ordering over a choice set a can be presented as follows. By a *preference ordering* $\succeq$ on a finite set A, we mean a binary relation, such that $x \succeq y$ may be either true or false for various pairs $x, y \in A$. When $x \succeq y$, we say "x is weakly preferred to y" (Kreps, 1990). We say $\succeq$ is *complete* if, for any $x, y \in A$, either $x \succeq y$ or $y \succeq x$. We say $\succeq$ is *transitive* if, for all $x, y, z \in A, x \succeq y$ and $y \succeq z$ imply $x \succeq z$. When these two conditions are satisfied, we say $\succeq$ is a *preference relation.* We say an agent *maximizes* $\succeq$ if, if from any subset $B \in A$, the agent chooses one of the most preferred elements of B according to $\succeq$,

> **Theorem**: *If $\succeq$ is a preference relation on set A, and if an agent maximizes $\succeq$, then there always exists a utility function u: Ato $\boldsymbol{R}$ (where $\boldsymbol{R}$ are the real numbers) such that the agent behaves as if maximizing this utility function over A.*

The empirical evidence supports an even stronger notion of human rationality for such preferences as charitable giving and punitive retribution. Andreoni and Miller (2002) have shown that one can apply standard choice theory, including the derivation of demand curves, plotting concave indifference curves, and finding price elasticities, in situations where individuals are faced with trade-offs between self-regarding and other-regarding payoffs. This is because individual preferences tend to satisfy the *Generalized Axiom of Revealed Preference*, which can be defined as follows. Suppose an agent chooses a commodity bundle $x_1, \ldots, x_n$ at prices $p_1, \ldots, p_n$ subject to the budget constraint $\sum_i p_i x_i = M$. Suppose $x^1, \ldots, x^m$ are any commodity bundles, so $x^j = (x_1^j, \ldots, x_n^j)$ for any $j = 1, \ldots, m$. Thus, x^j lies on the budget constraint if $\sum_i p_i x_i^j = M$. We say x^i is *directly revealed preferred* to x^j if x^j was in the choice set when x^i was chosen. We say x^1 is *indirectly revealed preferred* to x^n if there is some choice of $x^2, \ldots, x^{n-1}$ such that x^i is directly revealed preferred to x^{i+1} for $i = 1, \ldots, n-1$. Finally, we say that the Generalized Axiom of Revealed Preference (GARP) is satisfied if, whenever x^i is indirectly preferred to x^j, then x^i violates the income constraint when x^j is chosen.

Andreoni and Miller (2002) used a modified version of the dictator game, in which the experimenter gives a subject an amount of money, with the instructions that he is to share the money with a second party, specified by the experimenter, in any proportions that he wishes. The recipient has no say in the matter. In the current experiment, the subject was given an amount of money m, of which he could keep an amount p_s of his choosing, the remainder, $m - p_s$, being divided by the 'price' p and given to the second party. It is easy to see that the 'commodity bundle' (π_s, π_o) satisfies the budget equation $\pi_s + p\pi_o = m$.

The shape of the subject's preference ordering, and in particular whether it satisfies GARP, could be determined by varying the price p and the income m, and observing the subject's choices.

The experimenters found that 75% of subjects exhibited some degree of other-regarding preferences (i.e., gave money to the second party), and 98% of subjects made choices compatible with GARP. In some of the cases, p was chosen to be negative over some range, within which subjects maximize their own payoff by contributing *more* to the second party. Even in these cases GARP was generally satisfied, 23% of subjects exhibiting *jealous* preferences, by making a non-personal-payoff-maximizing choice, the sole attraction of which is that it reduces the payment to the second party.

While much more experimentation of this sort remains to be carried out, at least at this point it appears that other-regarding preferences present no challenge to traditional consumer theory.

3. Backward Induction and Rationality

Game theory privileges subgame perfection as the proper equilibrium concept of rational agents (Selten 1975). Subgame perfection, of course, is equivalent to the iterated elimination of weakly dominated strategies. It has long been known, however, that subjects in experimental games rarely engage in more that a few iterations of backward induction. In his ambitious overview of the current state of behavioral game theory Camerer (2003) summarizes a large body of experimental evidence in the following way: "Nearly all people use one step of iterated dominance ... However, at least 10% of players seem to use each of two to four levels of iterated dominance, and the median number of steps of iterated dominance is two." (202)

In this section, I will outline the empirical basis for this assertion. Despite its importance, I want to stress that this empirical regularity does not in any way undermine the rational actor model, since the interactive decision theoretic literature clearly shows that strong informational assumptions are necessary to justify the iterated elimination of (weakly or strongly) dominated strategies.

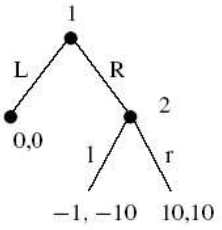
So-called 'beauty contests' are often used to determine the extent to which people backward induct. Suppose a group of subjects is told each should chose a whole number between zero and 100. The prize is \$10 and the winner is the subject whose guess is closest to 2/3 of the average guess. One level of backward induction implies limiting one's choice to [0,67], since this is the greatest 2/3 of the average can be. But, if everyone uses one level of backward induction, a subject knows that the highest average can be is 2/3 of 67, or about 44. With three levels of backward induction, the highest bid can be is 29, and with four levels, 20. If all players backward induct all the way, we get to the unique Nash equilibrium of zero.

Nagel (1995) was the first to study this beauty contest, using a group of fourteen to sixteen subjects. She found the empirical results to be compatible with the assertion that 13% of subjects used no backward induction, 44% used

one level, 37% used two levels and less than 4% use more that two levels of backward induction.

The failure of individuals to use backward induction is quite a shocker for classical game theorists, who tend to consider eliminating weakly dominated strategies a key element of rationality. However, assuming *common knowledge of rationality along the game path*, which means agents maximize their payoffs whenever it is their turn to play, given their conjectures as to the behavior of the other players, Bernheim (1984) and Pearce (1984) showed agents will only obey the iterated elimination of *strictly* dominated strategies. Such strategies are termed *rationalizable*. In a game like the centipede game or the finitely repeated prisoner's dilemma, there are no strictly dominated strategies, so *any* strategy is rationalizable. If we were to accept rationalizability as the criterion of rationality, the observation that people engage in limited backward induction would not entail their irrationality.

Unfortunately, the concept of rationalizability embodies an excessively weak concept of rationality, since it assumes nothing concerning the behavior of agents *off* the path of play. If a player moves at more than a single information set, backward induction in general eliminates *weakly* dominated strategies, so it is clear that even the simplest sort of incredible threats are rationalizable. Consider, for instance, the extensive form game in Figure 1. It is clear that a rational Player 2 will not play l if he gets to move, but the normal form of this game, shown to the right in this figure, indicates that l is only weakly dominated by r, so both l and r are rationalizable for player 2, from which it follows that both L and R are rationalizable for player 1.



	l	r
L	0,0	0,0
R	−1, −10	10,10

Figure 1: A Rationalizable Incredible Threat. The left pane is the extensive form game and the right pane its corresponding normal form game.

By contrast, Aumann (1995), following seminal contributions by Kripke (1959) and Lewis (1969), has shown that in generic games of perfect information (all information sets are singletons) if there is common knowledge of rationality both *on* and *off* the game tree, meaning that at each node, the player who moves acts to maximize his payoff in the subgame beginning at that node, then *only the backward induction solution remains*. To understand his argument, suppose there are players $i = 1, \ldots, n$ with strategy sets S_i for player i. Let Ω be the

set of *states of the world*. For illustration purposes, suppose the game is that of Figure 1, where the most natural choice is $\Omega = \{Ll, Lr, Rl, Rr\}$, corresponding to the possible 'types' of the two players, Ll corresponding to a 'left-type player one' and a 'left type player 2', and similarly for the remaining three states. A *knowledge partition* for a player is a partition of Ω into non-overlapping subsets $\mathcal{E}_i = E_i^1, \ldots, E_i^{k_i}$, called *events*, with the interpretation that if the actual state is $\omega \in E_i^j$, then i knows only that the actual state is somewhere in E_i^j. For instance, in our example, if each player knows only his own type, then $\mathcal{E}_1 = \{\{Ll, Lr\}, \{Rl, Rr\}\}$ and $\mathcal{E}_2 = \{\{Ll, Rl\}, \{Lr, Rr\}\}$. Note that we assume that whatever else may be in a state of the world, the moves of the various players are among them. Aumann formalizes this by assuming that a knowledge system includes a map $\mathbf{s} : \Omega \to \times_i S_i$ such that $\mathbf{s}(\omega) = (\mathbf{s}(\omega)_1, \ldots, \mathbf{s}(\omega)_n)$ is the strategy that each player chooses in state ω. The informational assumption is formalized by requiring that for any player i and any $E_i^j \in \mathcal{E}_i$ and any $\omega, \omega' \in E_i^j$, we have $\mathbf{s}(\omega)_i = \mathbf{s}(\omega')_i$; i.e., a player must make the same move at all events in one of his knowledge partition sets.

Suppose E is any event (i.e., any nonempty subset of Ω). K_iE denotes the union of all elements of $\mathcal{E}_i$ contained in E. We interpret K_iE as the event that i *knows* event E. We then write $KE = \cap_i K_iE$, which is the event that all players know event E. Finally, we write

$$CKE = KE \cap KKE \cap KKKE \cap \ldots$$

which is the event that E is *common knowledge*. For instance, in our example, the only event that is common knowledge is Ω itself. To see this, suppose K_1E includes $\{Ll, Lr\}$. Then K_2K_1E is either empty, or includes $\{Rl\}$ and $\{Rr\}$, in which case $E = \Omega$. A similar argument for the other partition elements shows that Ω is the only event that is common knowledge. We may also determine the event R_i that players i is rational, which is simply $R_1 = \{Ll, Rr\}$ and $R_2 = \{Lr, Rr\}$. Finally, we can identify the event that backward induction is used as being $I = \{Rr\}$.

Using this terminology, Aumann proves, under the stated conditions, that $CKR \subseteq I$, where R is the event that all players are rational, and I is the event that the backward induction solution is chosen. In the case of our example, there is no state at which there is common knowledge of rationality. However, let us expand the knowledge partitions to $\mathcal{E}_1 = \mathcal{E}_2 = \{Ll, Lr, Rl, Rr\}$, which says each player knows both his own and the other's type. Then $R = R_1 \cap R_2 = \{Rr\}$, so $CKR = KR = R = I$, which shows that common knowledge of rational implies backward induction.

There are good reasons, however, to explore alternatives to Aumann's treatment of common knowledge. It is easy to show that $K_iE \subseteq E$ for any i and any E, which means that to know an event implies that it is true (i.e., to know that you are in one of a set of states implies that you *are* in one of those states). But, rationality off the game tree by player i includes rationality at nodes where i moves, but that could only be reached by i having previously behaved non-rationally!

Consider, for instance, the centipede game, depicted in Figure 2. Backward

induction implies J will play D on the last round, awarding him 101 instead of 100. But to preempt this, M will play D on the next-to-last round, awarding him 101 instead of 98. Reasoning similarly, proceeding backward, we see that M will play D on the first round of play. Thus immediate defection (playing D) is the only subgame perfect Nash equilibrium to this game. Indeed, a little reflection will convince the reader that immediate defection is the only Nash equilibrium for both players.

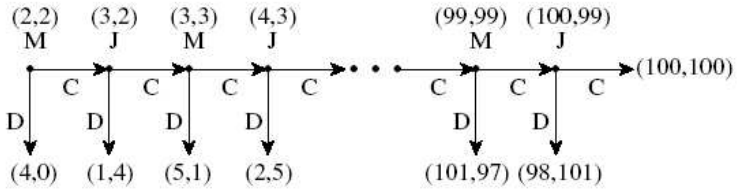


Figure 2: The Hundred Round Centipede Game

However, backward induction on the last round makes sense only if J is rational there. But, to get there J must have been *irrational* in each of his last 49 moves! The only way this makes any sense at all is if we assume that players can make mistakes, and that both players have made almost fifty mistakes in a row. This is hardly a plausible assumption on which to base a justification of backward induction.

An alternative model of rationality, allowing individuals to change their assessment of an opponent's type when he has made an unexpected move. One such has been developed by Ben-Porath (1997), which employs the concept of *certainty* in the place of *knowledge*, the difference being that *certainty* is a subjective state of assigning probability one to an event, even though the event may not contain the true state of the system. Rationality in this framework is defined as maximizing expected payoffs with a given set of expectations concerning the behavior of the other player. For instance, two players of the 100-stage centipede game might hold common certainty of rationality, which implies the backward induction solution if it holds at all nodes, but if player one cooperates on the first round, player two then drops his assumption in favor of some other subjective probability distribution over how his opponent will act at later nodes in the game tree.

Ben-Porath shows that common certainty of rationality is equivalent to one round of elimination of weakly dominated strategies, followed by any number of rounds of elimination of strictly dominated strategies. Thus, in Figure 2, only the final two decision nodes can be eliminated by common certainty of rationality. This makes a good deal of sense. Both M and J may be rational in the sense of attempting to maximize their expected payoffs, and by playing C on his first move, M signals that he is not playing backward induction. In deciding how far

to cooperate with M, J must have some probability distribution over where M will first defect, and choose a defection point that maximizes his payoff subject to this probability distribution. Experimental evidence (McKelvey/Palfrey 1992) indicates that subjects generally cooperate until the last few rounds.

Nevertheless, in some cases the concept of rationalizability fails to encompass the bounds of rational behavior. Consider, for instance the following version G_n of Kaushik Basu's Traveler's Dilemma (Basu 1994). Two business executives pay bridge tolls while on a trip, but cannot get receipts. Their superior tells each of them to report an integral number of dollars between \$2 and n on their expense sheet. If they report the same number, each will receive this much back. If they report different numbers, they each get the smaller amount, plus the low reporter gets an additional \$2, and the high reporter loses \$2. The executives are not permitted to collude in deciding what to report.

Let s_k be the strategy 'report k'. It is then easy to show that $n > 3, s_n$ in the game G_n is strictly dominated by a mixed strategy of $s_2, \ldots, s_{n-1}$. First, a glance at the normal form matrix for G_4 shows that s_4 is strictly dominated by a mixed strategy σ_4 using (i.e., weighting with positive probability) only s_2 and s_3. Second, it is easy to see that for any $n > 4$, if s_{n-1} is strictly dominated by a mixed strategy σ_{n-1} using only σ_{n-2} and s_2 in G_{n-1}, then s_n is strictly dominated by a mixed strategy σ_n using only σ_{n-1} and s_{n-1} in G_n. By the iterated elimination of strictly dominated strategies, this shows that the only rationalizable strategy of G_n is for both players to ask for \$2. This is also the only Nash equilibrium. Yet, it is clear that people will not generally play anything even approximating this equilibrium. Moreover, it is easy to see that this result does not depend on the size of the penalty! Any positive amount will do. In a beautiful experiment Capra et al. (1999), show that for small penalties, players in G_{100} play near 100, while for large penalties, they play near the rationalizable/Nash equilibrium of the game, which is $k = 2$.

I have barely scratched the surface in the modeling of rationality in the interactive decision theory literature. However, I have presented enough to make it clear that the empirical evidence on limited nature of backward induction exhibited by human subjects does not call into question the rationality of human subjects.

4. Complete Markets and the Self-Regarding Preferences Model

Contemporary economic theory holds that in a market for a product, the equilibrium price is at the intersection of the supply and demand curves for the good. Indeed, it is easy to see that at any other point a self-regarding seller could gain by asking a higher price, or a self-regarding buyer could gain by offering a lower price. This situation was among the first to be simulated experimentally, *the prediction of market-clearing virtually always receiving strong support* (Holt 1995). Here is a particularly dramatic example, provided by Holt et al. (1986) (reported by Charles Holt in Kagel/Roth 1995).

In the Holt, Langan, and Villamil experiment there are four 'buyers' and four 'sellers'. The good is a chip that the seller can redeem for $5.70 but the buyer can redeem for $6.80 at the end of the game. In analyzing the game, we assume throughout that buyers and sellers are self-regarding. In each of the first five rounds, each buyer was informed, privately, that he could redeem up to four chips, while eleven chips were distributed to sellers (three sellers were given three chips each, and the fourth was given two chips). Clearly, buyers are willing to pay up to $6.80 per chip for up to four chips each, and buyers are willing to sell their chip for any amount at or above $5.70. Total demand is thus sixteen for all prices at or below $6.80, and total supply is eleven chips at or above $5.70. Since there is an excess demand for chips at every price between $5.70 and $6.80, the only point of intersection of demand and supply curves is at the price *p=$6.80.* The subjects in the game, however, have absolutely no knowledge of aggregate demand and supply, since each knew only his own supply of or demand for chips.

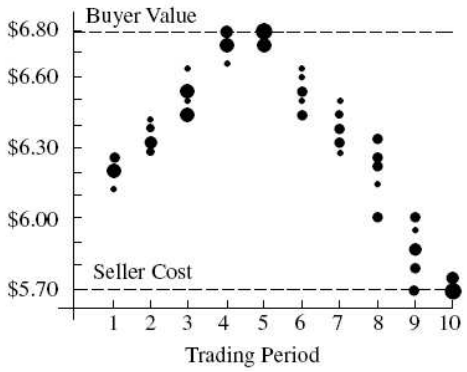


Figure 3: Simulating a Market Equilibrium: The Double Auction. The size of the circle is proportional to the number of trades that occurred at the stated price.

The rules of the game are that at any time a seller can call out an asking price for a chip, and a buyer can call out an offer price for a chip. This price remains 'on the table' until either it is accepted by another player, or a lower asking price is called out, or a higher offer price is called out. When a deal is made, the result is recorded and that chip is removed from the game. As seen in Figure 3, in the first period of play, actual prices were about midway between $5.70 and $6.80. Over the succeeding four rounds the average price increased, until in period 5, prices were very close to the equilibrium price predicted by traditional price theory.

In period six and each of the succeeding four periods, buyers were given the right to redeem a total of eleven chips, and each seller was given four chips. In this new situation, it is clear (to us) that there is an excess supply of chips

at each price between \$5.70 and \$6.80, so the only place supply and demand intersect is at \$5.70. While sellers, who previously made a profit of about \$1.10 per chip in each period, must have been delighted with their additional supply, succeeding periods witnessed a steady fall in price, until in the tenth period, the price is close to the theoretical prediction, and now the buyers are earning about \$1.70 per chip. A more remarkable vindication of the standard market model would be difficult to imagine.

5. Strong Reciprocity in the Labor Market

Akerlof (1982) suggested that many puzzling facts about labor markets could be better understood if it were recognized that in many situations, employers pay their employees higher wages than necessary, in the expectation that workers will respond by providing higher effort than necessary. Fehr et al. (1997) performed an experiment to validate this *gift exchange* model of the labor market.

The experimenters divided a group of 141 subjects (college students who had agreed to participate in order to earn money) into 'employers' and 'employees'. The rules of the game are as follows. If an employer hires an employee who provides effort e and receives a wage w, his profit is $p = 100e - w$. The wage must be between 1 and 100, and the effort is between 0.1 and 1. The payoff to the employee is then $u = w - c(e)$, where $c(e)$ is the 'cost of effort' function shown in Figure 4. All payoffs involve real money that the subjects are paid at the end of the experimental session. We call this the *experimental labor market game*.

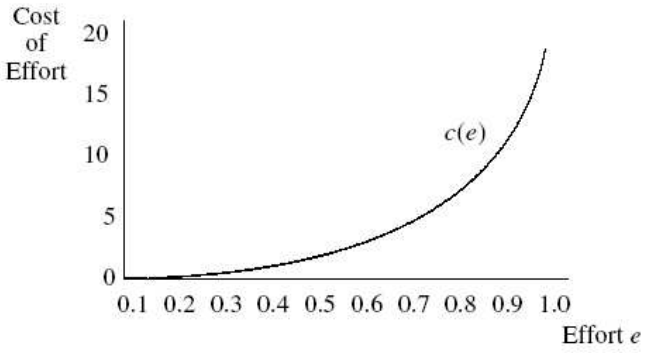


Figure 4: The Cost of Effort Schedule in Fehr, Gächter, and Kirchsteiger (1997).

The sequence of actions is as follows. The employer first offers a 'contract' specifying a wage w and a desired amount of effort e^*. A contract is made with the first employee who agrees to these terms. An employer can make a

contract (w, e^*) with at most one employee. The employee who agrees to these terms receives the wage w and supplies an effort level e, which *need not equal the contracted effort*, e^*. In effect, there is no penalty if the employee does not keep his promise, so the employee can choose any effort level, $e \in [0.1, 1]$, with impunity. Although subjects may play this game several times with different partners, each employer-employee interaction is a one-shot (non-repeated) event. Moreover, the identity of the interacting partners is never revealed.

If employees are self-regarding, they will choose the zero-cost effort level, $e = 0.1$, no matter what wage is offered them. Knowing this, employers will never pay more than the minimum necessary to get the employee to accept a contract, which is 1 (assuming only integral wage offers are permitted).[2] The employee will accept this offer, and will set $e = 0.1$. Since $c(0.1) = 0$, the employee's payoff is $u = 1$. The employer's payoff is $p = 0.1 \times 100 - 1 = 9$.

In fact, however, this self-regarding outcome rarely occurred in this experiment. The average net payoff to employees was $u = 35$, and the more generous the employer's wage offer to the employee, the higher the effort provided. In effect, employers presumed the strong reciprocity predispositions of the employees, making quite generous wage offers and receiving higher effort, as a means to increase both their own and the employee's payoff, as depicted in Figure 5. Similar results have been observed in Fehr, Kirchsteiger and Riedl (1993, 1998).

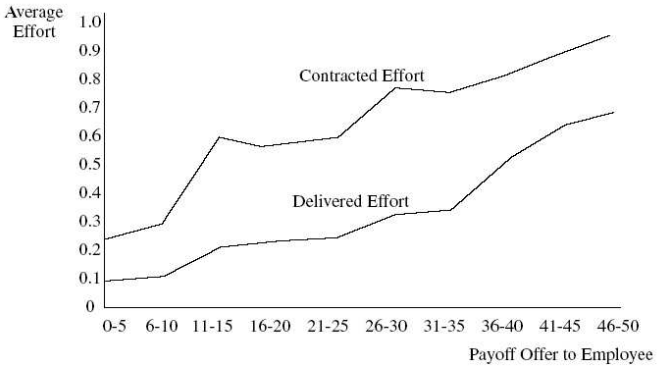


Figure 5: Relation of Contracted and Delivered Effort to Worker Payoff (141 subjects). From Fehr, Gächter, and Kirchsteiger (1997).

Figure 5 also shows that, though most employees are strong reciprocators, at any wage rate there still is a significant gap between the amount of effort agreed upon and the amount actually delivered. This is not because there are a few 'bad apples' among the set of employees, but because only 26% of employees

[2] This is because the experimenters created more employees than employers, thus ensuring an excess supply of employees.

delivered the level of effort they promised! We conclude that strong reciprocators are inclined to compromise their morality to some extent.

To see if employers are also strong reciprocators, the authors extended the game by allowing the employers to respond reciprocally to the *actual effort choices* of their workers. At a cost of 1, an employer could *increase* or *decrease* his employee's payoff by 2.5. If employers were self-regarding, they would of course do neither, since they would not (knowingly) interact with the same worker a second time. However, 68% of the time, employers punished employees that did not fulfill their contracts, and 70% of the time, employers rewarded employees who overfulfilled their contracts. Indeed, employers rewarded 41% of employees who *exactly* fulfilled their contracts. Moreover, employees *expected* this behavior on the part of their employers, as shown by the fact that their effort levels *increased significantly* when their bosses gained the power to punish and reward them. Underfulfilling contracts dropped from 83% to 26% of the exchanges, and overfulfilled contracts rose from 3% to 38% of the total. Finally, allowing employers to reward and punish led to a 40% increase in the net payoffs to all subjects, even when the payoff reductions resulting from employer punishment of employees are taken into account.

We conclude from this study that the subjects who assume the role of 'employee' conform to internalized standards of reciprocity, even when they are certain there are no material repercussions from behaving in a self-regarding manner. Moreover, subjects who assume the role of employer expect this behavior and are rewarded for acting accordingly. Finally, employers reward good and punish bad behavior when they are allowed, and employees expect this behavior and adjust their own effort levels accordingly. In general, then subjects follow an internalized norm not only because it is prudent or useful to do so, or because they will suffer some material loss if they do not, but rather because they desire to do so *for its own sake.*

6. Markets with Incomplete Contracting: The Economy as Social System

"An economic transaction", says Abba Lerner (1972), "is a solved political problem. Economics has gained the title of queen of the social sciences by choosing solved political problems as its domain." Lerner's observation is correct, however, only insofar as economic transactions are indeed *solved* political problems. The assumption in contemporary economic theory that gives this result is that *all economic transactions involve contractual agreements that are enforced by third parties (e.g., the judiciary) at no cost to the exchanging parties.* However, some of the most important economic transactions are characterized by the *absence of third-party enforcement.*

Consider, for instance, the relationship between a employer and an employee. The employer promises to pay the worker, and the worker agrees to work hard on behalf of the firm. The worker's promise, however, is typically not suitably specific to be enforceable in a court of law. Rather than suing an employee

for not working sufficiently hard, the employer generally simply dismisses the worker. For the threat of dismissal to be effective, the employer must pay a wage sufficiently high that the worker can expect incur substantial unemployment and search costs to secure an equally good alternative position. Hence, the exchange between employer and employee is not a 'solved political problem', and both the gift exchange issue analyzed in Section 5 and the disciplining of labor by virtue of the authority relationship between employer and employee may be involved in the determination of wages, labor productivity, and indeed the overall organization of the production process.

An experiment conducted by Brown et al. (2004) shows clearly that if third party enforcement is ruled out, employers prefer to establish long-term relationships with employees, offering a high wages, and using the threat of ending the relationship to induce high effort. Rather than market clearing determining the wage, as in standard labor market models, the result in this experiment is a labor market dominated by long-term relationships, with a positive level of unemployment in equilibrium, and employed workers enjoying a payoff advantage over unemployed workers. Labor market competition has little effect on the wage rate in this case, because employers will not rupture long-term relationships by hiring the unemployed at a lower wage.

Brown, Falk, and Fehr (BFF) used 15 trading periods with 238 subjects and three treatments. The first treatment was the standard complete contract condition (C condition) in which labor effort is contractually specified and guaranteed. The second treatment was an incomplete contract condition (ICF condition) with exactly the same characteristics, including costs and payoffs to employer and employee, as in Section 5. In addition, however, workers were given a payment of 5 points in each period that they were unemployed. In both conditions, subjects had identification numbers that allow long-term relationships to develop. The third treatment, which we call ICR, was identical to ICF, except that long-term relationships were ruled out (subjects received shuffled identification numbers in each experimental period). This treatment is thus identical to the gift exchange model in Section 5, except for the 5 point 'unemployment compensation'.

All contracts formally lasted only one period, so even long-term relationships had to be explicitly renewed in each period. If agents are self-regarding, it is easy to see that in the ICR treatment, all employees will supply the lowest possible effort $e = 1$, and employers will offer wage $w = 5$. Each firm then has a profit of $10e - 5 = 5$, and each worker has payoff $w - c(e) = 5 - c(0) = 5$. This outcome will also occur in the last period of the ICF treatment, and hence by backward induction, will hold in all periods. In the C treatment with self-regarding agents, it is easy to show that the employer will set $w = 23$ and require $e = 10$, so workers get $w - c(e) = 23 - c(10) = 5$ and employers get $10e - w = 100 - 23 = 77$ in each period. Workers are, in effect indifferent between being employed an unemployed in all cases.

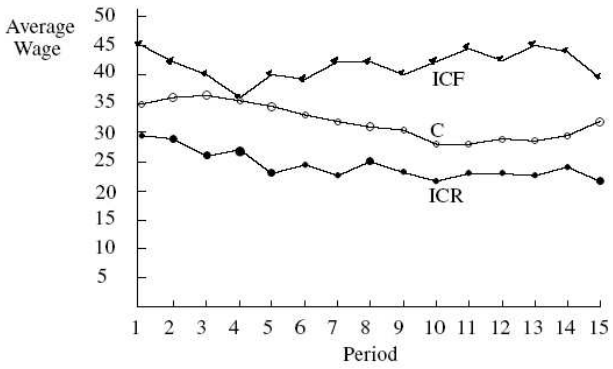


Figure 6: Wages over Fifteen Periods (Brown et al., 2004). The C treatment is complete contracting, the ICF treatment is incomplete contracting with long-term relationships permitted, and the ICR treatment is incomplete contracting with no long-term relationships permitted.

The actual results were, not surprisingly, quite at variance with the self-regarding preferences assumption. Figure 6 shows the path of wages over the fifteen periods under the three treatments. The ICR condition reproduces the result of Section 5, wages being consistently well above the self-regarding level of $w = 5$. If the C condition were a two-sided double auction, we would expect wages to converge to $w = 23$. In fact, the ICR conditions gives wages closer to the prediction for complete contracting than the C condition. The ICF condition gives the highest wages after the fourth period, validating the claim that under conditions of incomplete contracting, long-term relationships will prevail, and the distribution of gains will be more equal between buyers and sellers.

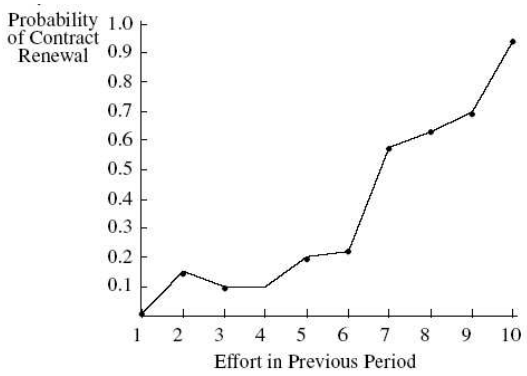


Figure 7: Contingent Renewal Provides Work Incentives in ICF Condition

By paying high wages in the ICF condition, employers were capable of effectively threatening their employees with dismissal (non-renewal of contract) if they were dissatisfied with worker performance. Figure 7 shows that this threat was in fact often exercised. Workers with effort close to $e = 10$ were non-renewed only about 5% of the time, whereas workers with effort below $e = 7$ were rarely renewed.

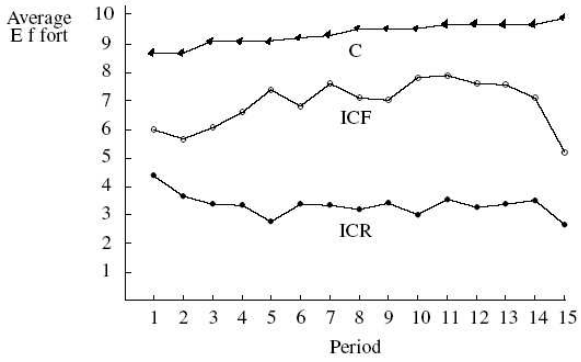


Figure 8: Worker Effort over Fifteen Periods. The C treatment is complete contracting, the ICF treatment is incomplete contracting with long-term relationships permitted, and the ICR treatment is incomplete contracting with no long-term relationships permitted.

Figure 8 shows that the effect of different contracting availabilities strongly affects the level of productivity of the system, as measured by average effort levels. Under complete contracting, effort levels quickly attain near-efficiency ($e = 10$), and remain there. Contingent renewal of long-term relationships achieves between 80% and 90% efficiency, with a significant end-game effect, as the threat of non-renewal is not very effective on the last few rounds. The gift exchange treatment (ICR), while supporting effort levels considerably above the self-regarding level, is considerably less efficient that either of the others, although it predictably suffers a smaller end-game effect that the ICF condition.

One extremely interesting pattern emerging from this study is the interaction of gift exchange and threat in the employer-employee relationship. One might think that they would be mutually exclusive, on the grounds that one cannot both feel charitable towards one's employer while at the same time being threatened by him. Yet, many of us will recall from personal experience this ambiguous co-presence of good will and fear. In this study, the importance of gift exchange in the long-term relationship is exhibited by the fact that even in the last two periods, where the threat of dismissal is weak or absent, effort levels are considerably above those of the pure gift exchange condition. Thus, gift exchange appears to be stronger when accompanied by the capacity of the

employer to harm, as though the fact that the employer has not exercised this capacity increases the worker's gratitude and willingness to supply effort.

7. Self-Regarding Explanations of Altruistic Behavior

The most general criticism of behavioral game theory is that the behavior of subjects in simple games under controlled and limited circumstances says nothing about their behavior in the extremely complex, rich, and temporally extended social relationships into which people enter in daily life. Defenders respond by reminding critics that controlled experiments have served the natural sciences extremely well, are very important in modeling animal behavior, as well as medical research and pharmacology, so it would be surprising if the ability to create conditions for controlled data collection were not equally valuable in the behavioral disciplines. Of course, it should be ascertained that the behaviors exhibited in pure form in the laboratory are operative as well in daily life. While there is much more to be done in this area, this appears to be the case, as shown in studies by Andreoni et al. (1998) on tax compliance. Bewley (2000) on fairness in wage setting, and Fong et al. (2005) on support for income redistribution, among others.

A second argument is that games in the laboratory are bizarre and unusual, so people really do not know how best to behave in these games. They therefore simply play as they would in daily life, in which interactions are repeated rather than one-shot, and take place among acquaintances rather than being anonymous. For instance, critics suggest that strong reciprocity is just a confused carryover into the laboratory of the subject's extensive experience with the value of building a reputation for honesty and willingness to punish defectors, both of which benefit the self-regarding actor. However, when opportunities for reputation building are incorporated into a game, subjects make predictable strategic adjustments compared to a series of one-shot games without reputation building, indicating that subjects are capable of distinguishing between the two settings (Fehr/Gächter 2000). Indeed, post-game interviews indicate that subjects clearly comprehend the one-shot aspect of the games. Moreover, subjects are often quite willing to punish others who do not harm them, but harm a third party by violating a social norm. It is not plausible to attribute this behavior to confusion with repeated games.

It is also simply not the case that we rarely face one-shot, anonymous interactions in daily life. Members of advanced market societies are engaged in one-shot games with very high frequency—virtually every interaction we have with strangers is of this form. Major rare events in people's lives (fending off an attacker, battling hand-to-hand in wartime, experiencing a natural disaster or major illness) are one-shots in which people appear to exhibit strong reciprocity much as in the laboratory. While members of the small-scale societies we describe below may have fewer interactions with strangers, they are no less subject to one-shots for the other reasons mentioned. Moreover, in these soci-

eties, greater exposure to market exchange led to stronger, not weaker, deviations from self-regarding behavior (Henrich et al. 2004).

Another indication that the other-regarding behavior observed in the laboratory is not simply error on the part of the subjects is that when experimenters point out that subjects could have earned more money by behaving differently, the subjects generally respond that of course they knew that, but preferred to behave in an ethically or emotionally satisfying manner rather than simply maximize their material gain.

An additional objection often expressed is that subjects really do not believe that the conditions of anonymity will be respected, and they behave altruistically because they fear their selfish behavior will be revealed to others. There are several problems with this argument. First, one of the strict rules of behavioral game research is that *subjects are never told untruths or otherwise misled*, and they are generally informed of this fact by experimenters. Thus, revealing the identity of participants would be a violation of scientific integrity. Second, there are generally no penalties for a self-regarding subject that could be attached to being 'discovered' behaving in a selfish manner—an other-regarding subject might feel shame, for example, but that is small comfort for the selfish actor model. Third, an exaggerated fear of being discovered cheating is itself a part of the strong reciprocity syndrome—it is a psychological characteristic that induces us to behave prosocially even when we are most attentive to our selfish needs. For instance, subjects might feel embarrassed and humiliated were their behavior revealed, but shame and embarrassment are themselves *other-regarding emotions* that contribute to prosocial behavior in humans (Bowles/Gintis 2003). In short, the tendency of subjects to overestimate the probability of detection and the costs of being detected are prosocial mental processes (H. L. Mencken once defined "conscience" as "the inner voice that warns us that someone may be looking"). Fourth, and perhaps most telling, in tightly controlled experiments designed to test the hypothesis that subject-experimenter anonymity is important in fostering altruistic behavior, it is found that subjects behave similarly regardless of the experimenter's knowledge of their behavior (Bolton/Zwick 1995; Bolton et al. 1998).

A final argument is that while a game may be one-shot and the players may be anonymous to one another, one will nonetheless *remember* how one played a game, and one may derive great pleasure from recalling one's generosity, or one's willingness to incur the costs of punishing another player for being selfish. This is quite correct, and probably explains a good deal of non-self-regarding behavior in experimental games.[3] But, this does contradict the fact that our behavior is other-regarding! Indeed, it confirms it, although there may be some philosophical arguments (irrelevant from the behavioral standpoint) that the other-regarding behavior is nonetheless self-interested.

[3] William Shakespeare understood this well when he has Henry V use the following words to urge his soldiers to fight for victory against a much larger French army: "Whoever lives past today ... will rouse himself every year on this day, show his neighbor his scars, and tell embellished stories of all their great feats of battle. These stories he will teach his son and from this day until the end of the world we shall be remembered."

Why, then, do individuals have other-regarding preferences? I believe there are three complementary reasons. The first is one that humans share very intimately with the rest of God's creatures: *kin altruism*. As laid out formally by William Hamilton (1963; 1964; 1970), it is directly fitness maximizing to incur a fitness cost c to afford a relative a fitness benefit b, provided $br \geq c$, where r is the degree of relatedness between the two. The second and third reasons are much more subtle, and if they have counterparts in the non-human world, they are almost certainly limited to primate species close to us on the evolutionary tree.

The first of these characteristically human mental structures are the *prosocial emotions*, including empathy, sympathy, shame, pride, and spite. These emotions are human universals (Brown 1991) that lead us to value the well-being of others as ends in themselves, in addition to whatever may be their contribution to our personal fitness and well-being. These emotions represent genetic predispositions in the sense that most humans are predisposed to exhibit them under the appropriate conditions, but what is considered 'appropriate' varies widely across different societies (Henrich et al. 2004; Bowles/Gintis 2003). In this sense, people contribute to cooperative endeavors even when this entails personal costs because it makes them feel good. What can be simpler? Moreover, the more costly it is to contribute, the more likely they are to shift their behavior towards activities with higher personal payoffs, be the self- or other-regarding.

The simple fact that it feels good to punish defectors was demonstrated clearly by de Quervain et al. (2004). The experimenters used PET scans of subject engaged in a trust game in which subjects could punish defection either symbolically or monetarily. They found that monetary punishment, as compared with symbolic punishment, activated brain regions that process the rewards that accrue as a result of goal-directed actions. Moreover, subjects with stronger activations in this area punished more vigorously.

The second predominantly human brain mechanism promoting other-regarding preferences is the *internalization of norms*, which involves older generations of well-socialized individuals (e.g., parents) and cultural institutions (e.g., schools, churches, tribal rituals) molding the preferences, norms, and values of youth in directions they deem desirable. An individual treats an internalized norm or value just as another argument in his preference function, and will devote costly resources towards meeting their normative requirements, employing the same adjudication mechanisms that are deployed in choosing among self-regarding goals (Andreoni/Miller 2002). The norm of reciprocity is among those that new members of society are socialized in every well-functioning society, although the content of the norm (i.e., what is considered fair and what is considered just punishment) vary widely across societies (Henrich et al. 2004).

8. Conclusion

Students of the history of science are aware of the centrality of *scientific instrumentation* in the progress of scientific knowledge. The microscope, the telescope,

electrophoresis, and a myriad of other tools of observation and data collection have laid the basis for our current understanding of the natural world. Each new tool allows for the construction of more subtle theoretical models, because it increases the power of observation to choose among models.

Behavioral game theory is simply one among a number of new scientific techniques that allow us to build better models of human behavior, and to move the discourse concerning human nature from the realm of political philosophy, where there has been little progress since the Eighteenth century, to the laboratory and the field, where stunning progress has, and doubtless will continue to be made. It is not too much to suggest that, with the addition of tools like behavioral game theory, neuroscientific instruments for assessing brain function, and agent-based computer simulation of life processes, the behavioral sciences may one day be put on the footing now enjoyed by the natural sciences.[4]

Bibliography

Ahlbrecht, M./M. Weber (1995), Hyperbolic Discounting Models in Prescriptive Theory of Intertemporal Choice, in: *Zeitschrift für Wirtschafts- und Sozialwissenschaften 115*, 535–568

Ainslie, G./N. Haslam (1992), Hyperbolic Discounting, in: G. Loewenstein/J. Elster (eds.), *Choice Over Time*, New York

Akerlof, G. A. (1982), Labor Contracts as Partial Gift Exchange, in: *Quarterly Journal of Economics 97(4)*, 543–569

Alcock, J. (1993), *Animal Behavior: An Evolutionary Approach*, Sunderland/MA

Allais, M. (1953), Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'école Américaine, in: *Econometrica 21*, 503–546

Andreoni, J./J. H. Miller (2002), Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism, in: *Econometrica 70(2)*, 737–753

— /B. Erard/J. Feinstein (1998), Tax Compliance, in: *Journal of Economic Literature 36(2)*, 818–860

Aumann, R. J. (1995), Backward Induction and Common Knowledge of Rationality, in: *Games and Economic Behavior 8*, 6–19

— /A. Brandenburger (1995), Epistemic Conditions for Nash Equilibrium, in: *Econometrica 65(5)*, 1161–80

Basu, K. (1994), The Traveler's Dilemma: Paradoxes of Rationality in Game Theory, in: *American Economic Review 84(2)*, 391–395

Ben-Porath, E. (1997), Rationality, Nath Equilibrium and Backward Induction in Perfect-Information Games, in: *Review of Economic Studies 64*, 23–46

Berg, J./J. Dickhaut/K. McCabe (1995), Trust, Reciprocity, and Social History, in: *Games and Economic Behavior 10*, 122–142

Bernheim, B. D. (1984), Rationalizable Strategic Behavior, in: *Econometrica 52(4)*, 1007–1028

Bewley, T. F. (2000), *Why Wages Don't Fall During a Recession,* Cambridge

Bolton, G. E./A. Ockenfels (2000), A Theory of Equity, Reciprocity and Competition, in: *American Economic Review 90(1)*, 166–193

[4] I would like to Samuel Bowles, Ernst Fehr, and Jack Hirshleifer for their help and the John D. and Catherine T. MacArthur Foundation for financial support. Affiliations: Santa Fe Institute, University of Siena, and Central European University

— /R. Zwick (1995), Anonymity versus Punishment in Ultimatum Games, in: *Games and Economic Behavior 10*, 95–121
— /E. Katok/R. Zwick (1998), Dictator Game Giving: Rules of Fairness versus Acts of Kindness, in: *International Journal of Game Theory 27(2)*, 269–299
Bowles, S./H. Gintis (1993), The Revenge of Homo Economicus: Contested Exchange and the Revival of Political Economy, in: *Journal of Economic Perspectives 7(1)*, 83–102
— / — (2003), The Origins of Human Cooperation, in: in P. Hammerstein (ed.), *The Genetic and Cultural Origins of Cooperation,* Cambridge/MA
Brown, D. E. (1991), *Human Universals*, New York
Brown, M./A. Falk/E. Fehr (2004), Relational Contracts and the Nature of Market Interactions, in: *Econometrica 72(3)*, 747–780
Camerer, C. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction,* Princeton/NJ
Capra, C. M./J. K. Goeree/R. Gomez/C. A. Holt (1999), Anomalous Behavior in a Traveler's Dilemma?, in: *American Economic Review 89(3)*, 678–690
Charness, G./M. Rabin (2002), Understanding Social Preferences with Simple Tests, in: *Quarterly Journal of Economics 117(3)*, 817–869
de Quervain, D. J.-F./U. Fischbacher/V. Treyer/M. Schellhammer/U. Schnyder/ A. Buck/E. Fehr (2004), The Neural Basis of Altruistic Punishment, in: *Science 305*, 1254–1258
di Finetti, B. (1974), *Theory of Probability*, Chichester
Ellsberg, D. (1961), Risk, Ambiguity, and the Savage Axioms, in: *Quarterly Journal of Economics 75*, 643–649
Fagin, R./J. Y. Halpern/Y. Moses/M. Y. Vardi (1995), *Reasoning about Knowledge,* Cambridge/MA
Fehr, E./S. Gächter (2000), Cooperation and Punishment, in: *American Economic Review 90(4)*, 980–994
— / — /G. Kirchsteiger (1997), Reciprocity as a Contract Enforcement Device: Experimental Evidence, in: *Econometrica 65(4)*, 833–860
— /G. Kirchsteiger/A. Riedl (1993), Does Fairness Prevent Market Clearing?, in: *Quarterly Journal of Economics 108(2)*, 437–459
— / — / — (1998), Gift Exchange and Reciprocity in Competitive Experimental Markets, in: *European Economic Review 42(1)*, 1–34
Fong, C. M./S. Bowles/H. Gintis (2005), Reciprocity and the Welfare State, in: H. Gintis/S. Bowles/R. Boyd/E. Fehr (eds.), *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life,* Cambridge/MA
Frey, B. (1997), A Constitution for Knaves Crowds Out Civic Virtue, in: *Economic Journal 107(443)*, 1043–1053
— (1997a), The Cost of Price Incentives: an Empirical Analysis of Motivation Crowding Out, in: *American Economic Review 87(4)*, 746–755
Gintis, H. (1976), The Nature of the Labor Exchange and the Theory of Capitalist Production, in: *Review of Radical Political Economics 8(2)*, 36–54
— (1989) The Power to Switch: On the Political Economy of Consumer Sovereignty, in: S. Bowles/R. C. Edwards/W. G. Shepherd (eds.), *Unconventional Wisdom: Essays in Honor of John Kenneth Galbraith,* New York, 65–80
— (2000), *Game Theory Evolving,* Princeton
Glimcher, P. W. (2003), *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*, Cambridge/MA

Güth, W./R. Schmittberger/B. Schwarz (1982), An Experimental Analysis of Ultimatum Bargaining, in: *Journal of Economic Behavior and Organization 3*, 367–388

Halpern, J. Y. (2001), Substantive Rationality and Backward Induction, in: *Games and Economic Behavior 37*, 425–435

Hamilton, W. D. (1963), The Evolution of Altruistic Behavior, in: *American Naturalist 96*, 354–356

— (1964), The Genetical Evolution of Social Behavior I & II, in: *Journal of Theoretical Biology 7*, 1–16, 17–52

— (1970), Selfish and Spiteful Behaviour in an Evolutionary Model, in: *Nature 228*, 218–1220

Hayashi, N./E. Ostrom/J. Walker/T. Yamagishi (1999), Reciprocity, Trust, and the Sense of Control: a Cross-societal Study, in: *Rationality and Society 11*, 27–46

Henrich, J./R. Boyd/S. Bowles/C. Camerer/E. Fehr/H. Gintis (2004), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-scale Societies*, Oxford

Holt, C. A. (1995), *Industrial Organization: A Survey of Laboratory Research,* Princeton

— /L. Langan/A. Villamil (1986), Market Power in an Oral Double Auction, in: *Economic Inquiry 24*, 107–123

Kagel, J. H./A. E. Roth (1995), *Handbook of Experimental Economics,* Princeton

Kahneman, D./A. Tversky (1979), Prospect Theory: An Analysis of Decision Under Risk, in: *Econometrica 47*, 263–291

Kreps, D. M. (1990), *A Course in Microeconomic Theory*, Princeton

Kripke, S. (1959), A Completeness Theorem in Modal Logic, in: *Journal of Symbolic Logic 24*, 1–14

Laibson, D. (1997), Golden Eggs and Hyperbolic Discounting, in: *Quarterly Journal of Economics 112(2,)* 443–477

Lerner, A. (1972), The Economics and Politics of Consumer Sovereignty, in: *American Economic Review 62(2)*, 258–266

Levine, D. K. (1998), Modeling Altruism and Spitefulness in Experiments, in: *Review of Economic Dynamics 1(3)*, 593–622

Lewis, D. (1969), *Conventions: A Philosophical Study*, Cambridge/MA

Machina, M. J. (1987), Choice under Uncertainty: Problems Solved and Unsolved, in: *Journal of Economic Perspectives 1(1)*, 121–154

Maynard Smith, J. (1982), *Evolution and the Theory of Games*, Cambridge

McKelvey, R. D./T. R. Palfrey (1992), An Experimental Study of the Centipede Game, in: *Econometrica 60*, 803–836

Nagel, R. (1995), Unravelling in Guessing Games: An Experimental Study, in: *American Economic Review 85*, 1313–1326

Pearce, D. (1984), Rationalizable Strategic Behavior and the Problem of Perfection, in: *Econometrica 52*, 1029–1050

Roth, A. (1995), Bargaining Experiments, in: J. Kagel/A. Roth (eds.), *The Handbook of Experimental Economics*, Princeton

Roth, A. E./V. Prasnikar/M. Okuno-Fujiwara/S. Zamir (1991), Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study, in: *American Economic Review 81(5)*, 1068–1095

Savage, L. J. (1954), *The Foundations of Statistics*, New York

Segal, U. (1987), The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach, in: *International Economic Review 28(1)*, 175–202

Selten, R. (1975), Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games, in: *International Journal of Game Theory 4*, 25–55

Shizgal, P. (1999), On the Neural Computation of Utility: Implications from Studies of Brain Stimulation Reward, in: D. Kahneman/E. Diener/N. Schwarz (eds.), *Well-Being: The Foundations of HedonicPsychology, N*ew York, 502–526

Sugden, R. (1993), An Axiomatic Foundation for Regret Theory, in: *Journal of Economic Theory 60(1)*, 59–180

Von Neumann, J./O. Morgenstern (1944), *Theory of Games and Economic Behavior,* Princeton

Watabe, M./S. Terai/N. Hayashi/T. Yamagishi Cooperation in the One-Shot Prisoner's Dilemma based on Expectations of Reciprocity, in: *Japanese Journal of Experimental Social Psychology 36*, 183–196

Yamagishi, T. (1986), The Provision of a Sanctioning System as a Public Good, in: *Journal of Personality and Social Psychology 51*, 110–116