Analyse & Kritik 27/2005 (© Lucius & Lucius, Stuttgart) p. 6–47

Ernst Fehr/Urs Fischbacher

Human Altruism—Proximate Patterns and Evolutionary Origins

Abstract: Are people selfish or altruistic? Throughout history this question has been answered on the basis of much introspection and little evidence. It has been at the heart of many controversial debates in politics, science, and philosophy. Some of the most fundamental questions concerning our evolutionary origins, our social relations, and the organization of society are centered around issues of altruism and selfishness. Experimental evidence indicates that human altruism is a powerful force and unique in the animal world. However, there is much individual heterogeneity and the interaction between altruists and selfish individuals is key for understanding the evolutionary dynamics as well as the proximate patterns of human cooperation. Depending on the environment, a minority of altruists can force a majority of selfish individuals to cooperate or, conversely, a few egoists can induce a large number of altruists to defect. Current gene-based evolutionary theories cannot explain important patterns of human altruism pointing towards the need for theories of cultural evolution and gene- culture coevolution.

Kin selection and . . . reciprocal altruism . . . are plausible as far as they go but I find that they do not begin to square up to the formidable challenge of explaining cultural evolution and the immense differences between human cultures around the world. . . . I think we have got to start again and go right back to first principles. For an understanding of the evolution of modern man we must begin by throwing out the gene as the sole basis of our ideas on evolution.

Richard Dawkins

Human societies represent a huge anomaly in the animal world. They are based on a detailed division of labour and cooperation of genetically unrelated individuals in large groups (Boyd and Richerson, 2005). This is obviously true for modern societies with their large organizations and nation states, but it also holds for hunter-gatherers which typically have dense networks of exchange relations and practice sophisticated forms of food-sharing, cooperative hunting, and collective warfare (Hill, 2002; Kaplan et al., 2000). In contrast, most animal species exhibit little division of labour and cooperation is limited to small groups. Even in other primate societies, with whom we share common ancestors, cooperation is orders of magnitude less developed than it is among humans. Exceptions are social insects such as ants and bees, or the naked mole rat; however, their cooperation is based on a substantial amount of genetic relatedness.

Why are humans such spectacular outliers with respect to all other animals? We propose that uniquely human forms of altruism provide the answer to this question. Human altruism goes far beyond that which has been observed in the animal world. Among animals, costly and fitness reducing acts which confer fitness benefits to other individuals are largely restricted to kin groups; despite several decades of research, evi-

dence for reciprocal altruism in pair-wise repeated encounters (Axelrod and Hamilton, 1981; Nowak and Sigmund, 1993; Trivers, 1971) remains scarce (Hammerstein, 2003). Likewise, there is little evidence that cooperation in non-human animals is affected by individual reputation building. This contrasts strongly with what we find in humans. If we randomly pick two human strangers from a modern society and give them the chance to engage in repeated anonymous exchanges in a laboratory experiment, reciprocally altruistic behaviour emerges spontaneously with a high probability (Andreoni and Miller, 1993; Gächter and Falk, 2002). However, human altruism even extends far beyond reciprocal altruism and reputation-based cooperation taking the form of strong reciprocity (Fehr et al., 2002b; Gintis, 2000). Strong reciprocity is a combination of altruistic rewarding – a readiness to reward others in response to fair outcomes or behaviour – and altruistic punishment – a willingness to sanction others for norm violations. Strong reciprocators bear the cost of rewarding or punishing but gain no individual economic benefits whatsoever from their acts. In contrast, reciprocal altruists, as they have been defined in the biological literature (Axelrod and Hamilton, 1981; Trivers, 1971), reward and punish only if this is in their long-term self-interest. Strong reciprocity thus constitutes a powerful incentive for cooperation even in non-repeated interactions and when reputation gains are absent because those who cooperate will be rewarded while those who defect will be punished by strong reciprocators.

The first part of this review is devoted to the experimental evidence documenting the relative importance of repeated encounters, reputation formation, and strong reciprocity for human cooperation. We do not discuss the role of kinship in human altruism because it is well known that humans share kin-driven altruism with many other animals (Daly and Wilson, 1988; Silk and 1980., 1980). We will show that the interaction between selfish and strongly reciprocal individuals is decisive for the understanding of human cooperation. We identify conditions under which selfish individuals trigger the breakdown of cooperation, and conditions under which the strongly reciprocal individuals have the power to ensure wide-spread cooperation. Next we discuss the limits of human altruism limits arising from the costs of altruistic acts, from competition among individuals, and from group boundaries. Then we show how recently developed mathematical theories of human motivation provide powerful tools for understanding and predicting proximate patterns of altruistic behaviours. Finally, we discuss the evolutionary origins of the different forms of human altruism. We are particularly interested in whether strong reciprocity represents an adaptive trait that can be explained by recent evolutionary models, whether these models can explain why humans, but not other animals, exhibit large-scale cooperation among genetically unrelated individuals, and to what extent the evidence supports the key ingredients of these models.

1. Proximate patterns

Suppose you observe that a friend of yours incurs cost to help another person. Why does your friend do this? He might be concerned about his reputation with respect to you because you observe him. Alternatively, he might be concerned about his reputation vis à vis the person he assisted and expect that this person will help him in future encounters. A third possibility is that your friend is truly a nice person who even helps if no reputation is at stake and future encounters are highly unlikely. Although you may have an opinion about why your friend helped you can never be sure. Observations of behaviour in real life circumstances will almost never enable you to discriminate between the different motives for helping. This is the reason why sound knowledge about the motives

behind altruistic acts predominantly comes from laboratory experiments. In the laboratory, the researcher controls the anonymity conditions and the possibilities for future interactions and reputation formation, which enables him to discriminate between different motives for helping or punishing. Therefore, we first discuss experiments in which interactions among kin, repeated encounters, and reputation formation have been ruled out. In a sense, altruistic acts occurring in this environment are "truly" altruistic because no future economic benefits for the individual or the individual's genes are possible. Next, we document how the possibility of future encounters and individual reputation formation changes subjects' behaviour. In all experiments discussed below, real money, sometimes up to three months' income (Cameron, 1999; Fehr et al., 2002a; Hoffman et al., 1996; Slonim and Roth, 1998), was at stake. Subjects never knew the personal identities of those with whom they interacted and they had full knowledge about the structure of the experiment – the available sequence of actions and the prevailing information conditions. If, e.g., the experiment ruled out future encounters between the same individuals, subjects were fully informed about this. To rule out any kind of social pressure, the design of the experiment even ensured in several instances that the experimenter could not observe subjects' individual actions but only the statistical distribution of actions (Bolton et al., 1998; Bolton and Zwick, 1995; Hoffman et al., 1994).

1.1 Altruistic Punishment

The ultimatum game (UG) (Güth et al., 1982) nicely illustrates that a sizeable number of people from a wide variety of cultures (Henrich et al., 2001; Roth et al., 1991) facing high monetary stakes (Cameron, 1999; Hoffman et al., 1996; Slonim and Roth, 1998) are willing to hurt others to prevent unfair outcomes or to punish unfair behaviour. In the UG, two subjects have to agree on the division of a fixed sum of money. Person A, the proposer, can make exactly one proposal of how to divide the money. Then person B, the responder, can accept or reject the proposed division. In case of rejection, both receive nothing whereas in case of acceptance, the proposal is implemented. A robust result in this experiment is that proposals that give the responder shares below 25 percent of the available money are rejected with a very high probability. This shows that responders do not behave in a self-interest maximizing manner because a selfish responder would accept any positive share. In general, the motive indicated for the rejection of positive, yet "low", offers is that responders view them as unfair. Most proposers seem to understand that low offers will be rejected. Therefore, the equal split is often the modal offer in the UG (Güth et al., 1990). The decisive role of rejections in the UG is indicated by the dictator game in which the proposer unilaterally dictates the division of the money because the responder cannot reject the offer. In the dictator game, the modal amount given to the responders is often zero (Forsythe et al., 1994; Hoffman et al., 1994).

To what extent do rejections in the UG confer benefits on other people? The altruistic dimension of rejections emerges when there is a population of proposers and responders, who repeatedly play the UG in such a way that no pair of

players meets again and no player knows anything about the previous history of his opponent. If the proposers believe that there are many altruistic punishers among the responders, they have a reason for making high offers, which benefit all responders. Thus, rejections are individually costly acts, which contribute to the collective reputation of the responders for being tough bargainers. They are altruistic acts because every responder, i.e., also one who does not reject, benefits from the higher offers that are generated by the collective reputation. For the purpose of this review, we ran an experiment with 10 proposers who met a different responder in 10 successive rounds. We observed that proposers who experienced a rejection in the previous round increased their offers in the current round by 6.5 percent.

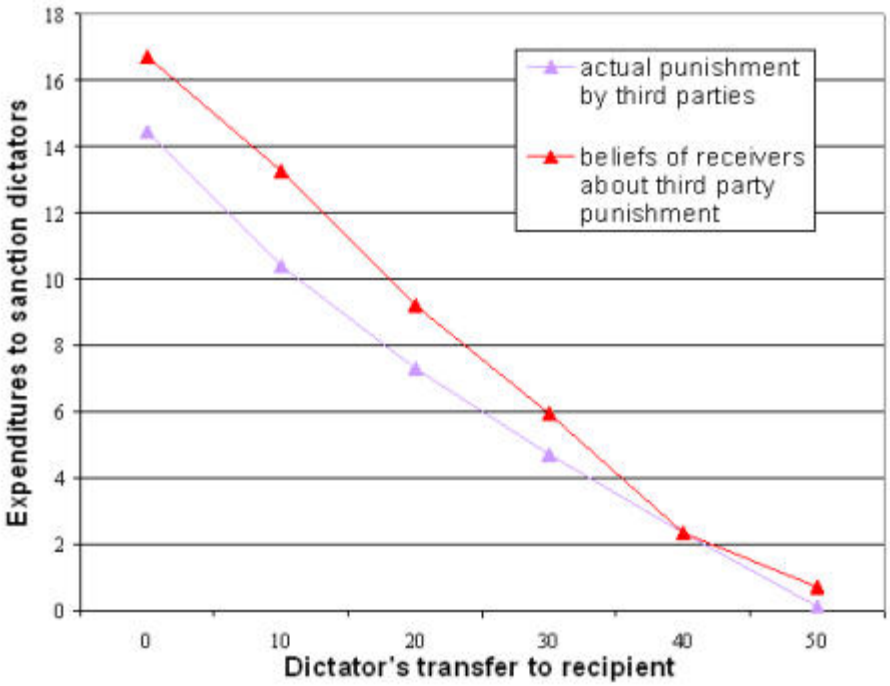


Figure 1: Altruistic punishment by third parties who are not directly affected by the violation of a fairness norm. The fair transfer level is given by 50. The more the dictator's transfer falls short of the fair level of 50, the more third parties punish the dictator. The recipients of the transfer also expect that the dictators will be punished for unfair transfers. Data are taken from (Fehr and Fischbacher, 2004)

In the UG, the responder is directly affected by the action of the proposer. However, a key element of the enforcement of many social norms, such as food-sharing norms in hunter-gatherer societies (Hill, 2002; Kaplan et al., 2000), is that people punish norm violators not for what they did to themselves but for

what they did to others (Bendor and Swistak, 2001; Sober and Wilson, 1998). Norm enforcement involves the punishment of norm violations even by those who are not economically affected by the violation. To study this question experimentally, we conducted a third-party punishment game (Fehr and Fischbacher, 2004). In the third-party punishment game (TPG) there are three subjects. Subject A is endowed with 100 money units (MUs), subject B has nothing and subject C is endowed with 50 MUs. A is in the role of a dictator, i.e. he is free to give whatever he wants to the "poor" subject B who is just a passive recipient of A's transfer. After the third party C has been informed about A's transfer to B he can spend money to punish A. Every MU spent on punishment reduces A's income by three MUs. The TPG is played exactly once and, due to anonymity, nobody can gain any reputation. Therefore, no selfish third party will ever spend money on punishment. Yet, we hypothesized that a fairness norm applies to this situation and, therefore, altruistic punishers were expected to punish A for sending unfairly low transfers to B. In fact, 55 percent of the third parties (n = 22) punish player A for transfers below 50 and punishment increases the lower the transfer (Figure 1). Moreover, between 70 and 80 percent of the recipients (n = 22) expect that dictators will be punished for transfers below 50. The recipients also expect that lower transfers will be punished more strongly, indicating that lower transfers are interpreted as more severe norm violations. Punishment by third parties creates an obvious benefit for the recipient because it deters dictators from keeping all the money. Recent research shows that altruistic punishment by third parties also applies to prisoner dilemmas (Fehr and Fischbacher, 2004): Third parties frequently punish a defector if the opponent of the defector cooperated. A further interesting regularity is the fact that, although third party punishment is quite frequent, it is less prevalent and less severe than altruistic punishment by parties affected directly.

1.2 Altruistic Rewarding

Sequentially played social dilemmas are a powerful tool for the study of altruistic rewarding. They come in various forms – as gift exchange games (Fehr et al., 1993), trust games (Berg et al., 1995), or sequentially played prisoners' dilemmas (Hayashi et al., 1999) – but the basic structure is captured by the following example. Two subjects, A and B, are endowed with, say, 10 MUs. Both subjects can keep whatever they like but they can also transfer part or all of their endowment to their opponent. The experimenter doubles any amount sent to the other subject so that, collectively, the two subjects are best off if they transfer their whole endowment: If both keep what they have, each subject earns 10 MUs, if both transfer their whole endowment each earns 20 MUs. From a selfish viewpoint, however, it is best to keep one's own endowment and to hope that the other subject transfers the whole endowment, yielding 30 MUs for oneself and 0 for the other. This experiment mimics the essence of a vast number of real life situations. Any kind of mutually beneficial exchange that takes place in the absence of enforceable contracts is characterized by a similar structure. In

these situations, both players are better off exchanging their goods and favours but there is also a strong temptation to cheat.

To study altruistic rewarding in the experiment above, one subject, say A, first makes a transfer. Then B is informed about A's transfer and decides how much to send to A. It is obvious that a selfish player B will send back nothing regardless how much he received. In fact, however, more than 50 percent of subjects in the role of B transfer money and frequently B's transfer increases A's transfer (Berg et al., 1995; Fehr et al., 1993). This becomes particularly transparent in the sequential prisoners' dilemma in which both subjects have only two available actions – to send nothing or to send everything. Here, if subject A sends nothing, virtually all B's respond with a transfer of zero while if A sends everything a sizeable percentage of B's – often more than 50 percent, sometimes approaching 90 percent – also send everything (Hayashi et al., 1999). Like altruistic punishment, the presence of altruistic rewarding has also been documented in many different countries (Buchan et al., 2002), in populations with varying demographic characteristics (Bellemare and Kröger, 2003), and under stake levels approaching 2-3 months' income (Fehr et al., 2002a).

1.3 Strong reciprocity and multilateral cooperation

A decisive feature of human cooperation in hunter-gatherer societies is that it is not restricted to bilateral interactions but also takes place in relatively large groups of dozens or several hundred individuals. To what extent does strong reciprocity contribute to cooperation in public goods situations involving larger groups of individuals? By definition, a public good can be consumed by every group member regardless of the member's contribution to the good. Therefore, each member has an incentive to free-ride on the contributions of others. This incentive can be neatly captured by the following one-shot experiment. Suppose each of $n > 2$ subjects is endowed with 10 MUs which can be privately kept or spent on a group project. Subjects decide simultaneously on how much they keep and how much they contribute. The experimenter doubles the total amount spent on the group project and distributes the proceeds of the doubled amount equally among the n members. This means that for every MU spent on the project each group member, including the contributing subject, earns $2/n$ MUs which is less than 1 because of $n > 2$. Yet, the contributing subject has costs of 1, meaning that a selfish subject never contributes anything to the project in a one-shot experiment. This holds, although it would be collectively rational to contribute everything because if all participants keep their endowments, they earn 10 MUs each, whereas if all contribute their endowments they earn 20 MUs a piece.

In such public goods experiments, altruistic rewarding implies that an individual's contributions increase due to the expected contributions from the other subjects. Subjects reward others if others are expected to raise their cooperation. If the experiment is played only once, subjects typically contribute between 40 and 60 percent of their endowment. There is also strong evidence that higher expectations about others' contributions induce individual subjects to contribute

more (Dawes, 1980; Fischbacher et al., 2001; Messick and Brewer, 1983; Yamagishi and Kiyonari, 2000). The interesting fact is, however, that cooperation is rarely stable and deteriorates to rather low levels if the game is played repeatedly (and anonymously) for 10 rounds (Andreoni, 1988; Isaac et al., 1985; Isaac et al., 1984; Ledyard, 1995). Early researchers inferred from this that if subjects gain experience with the situation, they learn to play what is in their best selfish interests. This interpretation has turned out to be untrue because if the same subjects are given the chance to start a new 10 round game, they start again with high cooperation rates (Andreoni, 1988).

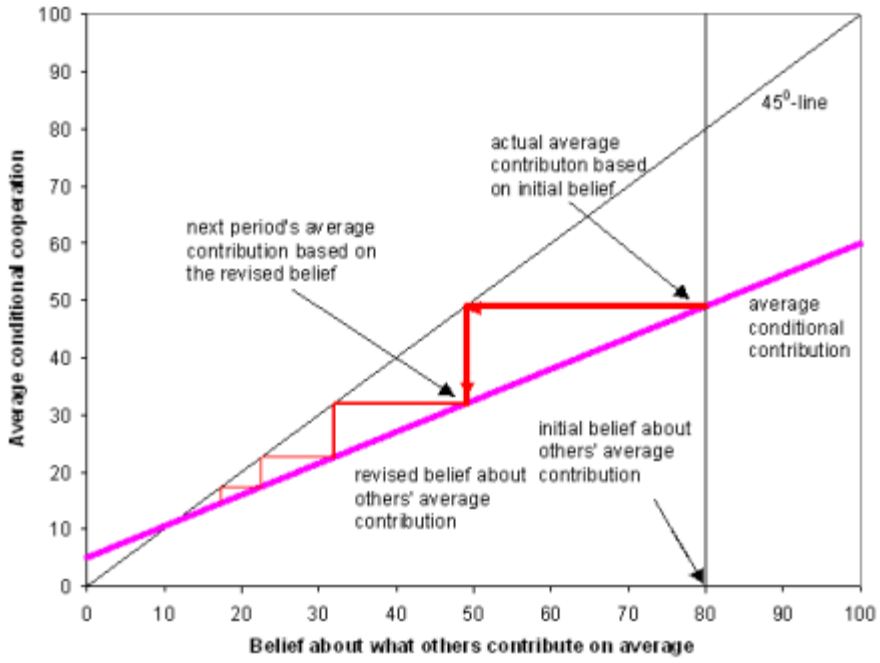


Figure 2: The decay of cooperation over time. Subjects are heterogenous with regard to their willingness to reward altruistically. This results in the relationship between the expected average contribution of other group members to the public good and the contribution of a representative individual (the average conditional cooperation) shown in the figure. Initially, individuals expect high average contribution rates, say 80 percent of the endowment. On average, this induces them to contribute 50 percent. Therefore, expectations are disappointed which leads to a downwards revision of expectations to say, 50 percent of the endowment. Yet, if individuals expect 50 percent they will in fact only contribute roughly 30 percent causing a further downwards revision of expectations. The process stops at the intersection point with the 45-degree line, which determines the equilibrium level of altruistic cooperation in this setting.

The most plausible interpretation of the decay of cooperation is based on the fact that subjects are heterogeneous in their degree of strong reciprocity. A recent study (Fischbacher et al., 2001) indicated that 10 percent of the subjects are willing to fully match the expected average contribution of other group members, 40 percent give somewhat less than the expected average contribution of others, 30 percent are complete free-riders who never contribute anything, 14 percent roughly match others' contribution until they spent half of their endowment, and the rest of the subjects exhibit quite irregular behaviour. Thus, on average, individual subjects increase their contribution levels in response to expected increases in the average contribution of other group members, but the intercept and the steepness of this relationship is insufficient to establish an equilibrium with high cooperation (Figure 2). In round 1, subjects typically have optimistic expectations about others' cooperation but, given the aggregate pattern of behaviours, this expectation will necessarily be disappointed, leading to a breakdown of cooperation over time.

This breakdown of cooperation provides an important lesson. Despite the fact that there are a large number of strong reciprocators, they cannot prevent the decay of cooperation under these circumstances. In fact, it can be shown theoretically that in a population with a clear majority of strong reciprocators, a small minority of selfish individuals suffices to render zero cooperation the unique equilibrium (Fehr and Schmidt, 1999). This implies that it is not possible to infer the absence of altruistic individuals from a situation in which we observe little cooperation. If strong reciprocators believe that no one else will cooperate, they also will not cooperate. To maintain cooperation in n-person interactions, the upholding of the belief that all or most members of the group will cooperate is thus decisive.

Any mechanism that generates such a belief has to provide cooperation incentives for the selfish and insufficiently reciprocal individuals. Direct punishment of non-cooperators in repeated interactions (Ostrom et al., 1992; Yamagishi, 1986) or altruistic punishment in one-shot interactions (Fehr and Gächter, 2002) provides one possibility. If cooperators have the chance to target their punishment directly towards those who defect and if the cost of sanctioning is lower than the cost of being sanctioned, punishment opportunities cause a large increase in cooperation levels and prevent a breakdown of cooperation (Fehr and Gächter, 2002; Ostrom et al., 1992; Yamagishi, 1986). The reason is that cooperators impose strong sanctions on the defectors, providing the necessary incentives for cooperation. Thus, in the presence of targeted punishment opportunities, strong reciprocators are capable of enforcing wide-spread cooperation by deterring potential non-cooperators. In fact, it can be shown theoretically that even a minority of strong reciprocators suffices to discipline a majority of selfish individuals when direct punishment is possible (Fehr and Schmidt, 1999).

1.4 Repeated interactions and reputation formation

A reputation for behaving altruistically is another powerful mechanism for the enforcement of cooperation in n-person public goods situations. If people are

engaged in bilateral encounters as well as in n-person public goods interactions, a defection in the public goods situation, if known by others, may decrease others' willingness to help in bilateral interactions. This idea has been neatly captured by the following experiment (Milinski et al., 2002b). Suppose that after each round of interaction in the public goods experiment, subjects play a so-called indirect reciprocity game (Nowak and Sigmund, 1998a, b). In this game subjects are matched in pairs and one subject is randomly placed in the role of a "donor" and the other in the role of a "recipient". The donor can help the recipient by giving an amount of money, say 2 MUs, which is doubled by the experimenter so that the recipient gets 4 MUs. The recipient's reputation is given by his decision in the previous public goods round and his history of helping decisions in the indirect reciprocity game. It turns out that the recipients' reputation in the public goods game is an important determinant for the donors' decisions. They punish the recipients by significantly reducing the likelihood of help when the recipients defected in the previous public goods game. This, in turn, has a powerful cooperation enhancing effect.

Helping behaviour in indirect reciprocity experiments has also been documented in the absence of interactions in public goods games (Milinski et al., 2001; Wedekind and Milinski, 2000). A crucial element in these experiments is that direct reciprocity is ruled out because no recipient will ever be put in a position where he can give to one of his previous donors. Helping rates between 50 and 90 percent can be achieved in these experiments, and recipients with a history of generous helping decisions are significantly more likely to receive help themselves. This suggests that the donors' behaviour may be driven by the desire to acquire a good reputation. However, it is also possible that altruistic rewarding drives helping behaviour. A neat way to study the relative importance of altruistic rewarding and reputation seeking in indirect reciprocity experiments is to allow only half the subjects in an experiment to acquire a reputation (Engelmann and Fischbacher, 2002). This means that one can compare the behaviour of donors who cannot gain a reputation with the behaviour of those who can. The data show that both altruistic rewarding and reputation seeking are powerful determinants of donors' behaviour. Donors without a reputation help in 37 percent of the cases whereas those with a reputation help in 74 percent of the cases. Moreover, donors with a reputation are much more likely to also help recipients with only a moderately generous history of helping whereas donors without a reputation target their help towards those recipients who have a high reputation for being generous.

These and other results (Milinski et al., 2002a; Seinen and Schram, in press) suggest that humans are very attentive to possibilities of individual reputation formation. They exhibit a sizeable baseline level of altruistic rewarding, and when given the opportunity to gain a reputation for being generous, helping rates strongly increase. Humans are similarly attentive to the possibility of repeated interactions with the same individual (reciprocal altruism). In sequential social dilemmas, the cooperation rate of the second movers is much higher if they know that there is a possibility of meeting the same partners again in future periods (Gächter and Falk, 2002). Likewise, in simultaneously played prisoners'

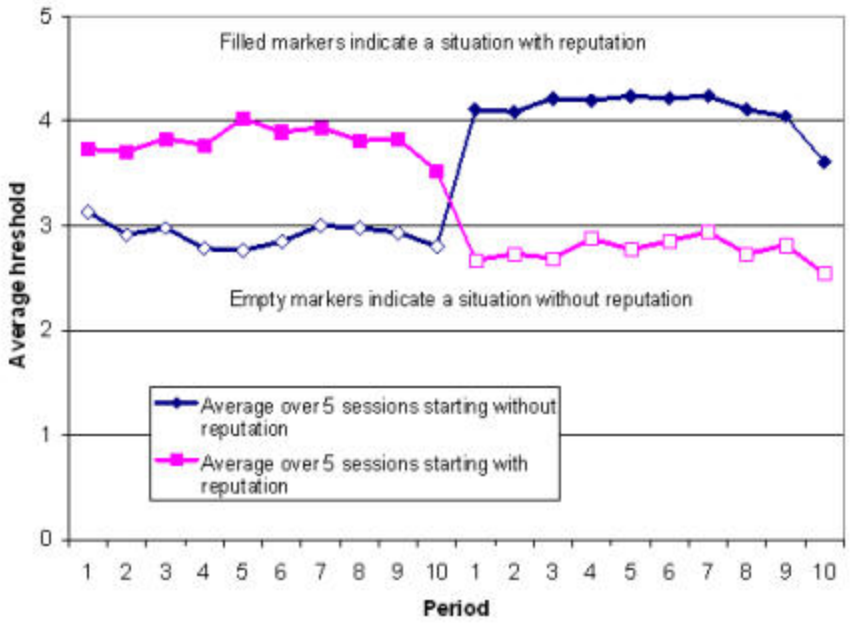
dilemmas, subjects are keenly attuned to the possibility of future interactions (Andreoni and Miller, 1993; DalBo, 2003). Cooperation rates strongly increase when the probability of future interactions with the same partner increases from zero to $\frac{1}{2}$ and further to $\frac{3}{4}$ (DalBo, 2003).

In the domain of rewarding behaviours, there is thus much evidence indicating that subjects understand the difference between one-shot and repeated encounters well, and also between interactions in which their reputation is or is not at stake. In the domain of punishing behaviours, however, little is known about repetition and reputation effects. For this purpose we conducted a computerized (Fischbacher, 1998) series of 10 ultimatum games in two conditions – a reputation condition and a baseline condition. In both conditions, there were 10 proposers and 10 responders and no proposer met the same responder more than once. In every period, the proposer suggests how to allocate 10 MUs between himself and the responder, who could accept or reject this proposal. In the reputation condition, the proposers were informed about the current responder's past rejection behaviour whereas in the baseline condition this knowledge was absent. This means that in the reputation condition, the responders could gain an individual reputation for being tough bargainers by rejecting high offers. A responder who incurred the short-term cost of a rejection could gain the long-term benefits of a "good" reputation by inducing future proposers to make him better offers. Since this economic benefit was absent in the baseline condition, subjects who understand the logic of reputation formation will exhibit higher acceptance standards in the reputation condition.

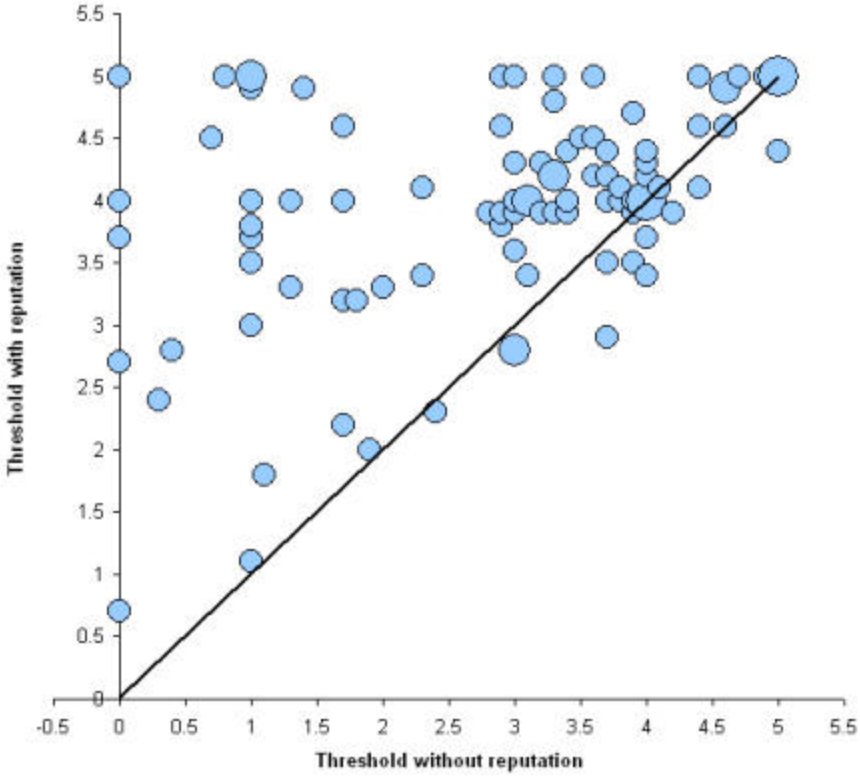
In both conditions, the responders indicated in each period an acceptance threshold before they knew the current offer. If the actual offer turned out to be above the threshold it was accepted, if it turned out to be below the threshold it was rejected. All subjects participated in a sequence of both conditions. The results show that when the subjects were first in the baseline condition, the average acceptance threshold was about 3 MUs whereas if they entered the reputation condition, their thresholds immediately jumped to more than 4 MUs (Figure 3a). This jump in the thresholds forced the proposers to increase their offers. Similarly, if the reputation condition took place first, the average thresholds immediately decreased when subjects entered the baseline condition (Figure 3a). Moreover, this change in the average thresholds is not an artefact of aggregation. It is explained by the fact that the vast majority of responders (82 percent, n = 94) increases the threshold in the reputation condition relative to the baseline (Figure 3b) while the remaining minority keeps the thresholds roughly constant. These results suggest that altruistic punishers clearly understand that, if individual reputation building is possible, it pays to acquire a reputation as a tough bargainer. This also means that their rejections in the baseline condition cannot be attributed to cognitive problems in understanding when individual reputation matters and when it does not matter.

1.5 Limits of human altruism

Strongly reciprocal individuals reward and punish in anonymous one-shot in-



(a) Responders' acceptance thresholds in the ultimatum game with and without reputation. **a**. Time trend of acceptance thresholds. If the control treatment without the opportunity to build an individual reputation for toughness is conducted first, the responders reject offers below 3 money units (blue line with empty markers). Immediately after the implementation of reputation building opportunities in period 11, the acceptance thresholds jump up to more than 4 money units indicating the desire to be known as a "tough" responder (blue line with filled markers). If the reputation treatment comes first (purple line) the removal of the opportunity to acquire a reputation immediately causes a decrease in responders' acceptance thresholds. Data are taken from (Fehr and Fischbacher, 2003)



(b) Individual level changes in responders' average acceptance thresholds. The relative size of the circles represents the frequency of observations behind a circle. Responders who increase their average acceptance threshold in the reputation condition relative to the baseline condition generate a data point above the 45-degree line. The vast majority of responders increase their thresholds when they can gain a reputation for toughness. Only a small minority lowers the thresholds or keeps them roughly constant. Data are taken from (Fehr and Fischbacher, 2003)

Figure 3:

teractions. Yet, they reward and punish more in repeated interactions or when their reputation is at stake. This suggests that they are motivated by a combination of altruistic and selfish concerns. Their altruistic motives induce them to cooperate and punish in one-shot interactions and their selfish motives induce them to increase rewarding and punishing in repeated interactions or when reputation building is possible. If this argument is correct, we should also observe that altruistic acts are less frequent if their costs are increased. At a higher cost, individuals have to give up more of their own payoff to help others, so that for a given combination of selfish and altruistic motives, the individuals will exhibit less altruistic behaviour. The evidence from dictator games and public good games confirms this prediction. If the own payoff that needs to be invested to produce one unit of the public good increases, subjects invest less into the public good (Isaac and Walker, 1988; Ledyard, 1995). Likewise, if in the dictator game the cost of transferring one MU to the recipient increases, the dictators give less money to the recipients (Andreoni and Miller, 2002).

Interestingly, if the cost of altruistic rewarding and punishing increases for others, subjects also anticipate that this will reduce others' strongly reciprocal behaviour. This has been shown in a sequential social dilemma with three stages (Fehr et al., 1997): First subject A decided how much to cooperate. Then subject B chose the cooperation level and, finally, subject A had the option to reward or punish B after he observed B's cooperation level. In the low cost condition, every MU invested by A into rewarding or punishing B increased or decreased B's payoff by 2.5 MUs. In the high cost condition, A's cost of rewarding and punishing were increased by a factor of 5. In the low cost condition, subjects in the role of A rewarded high and punished low cooperation by subject B much more than in the high cost condition. Moreover, subjects in the role of B anticipated the decrease in rewarding and punishment in the high-cost condition and, hence, cooperated much less. This suggests that humans have a well developed capability of judging the determinants of other people's altruistic behaviour.

A particularly interesting form of cost change occurs if one introduces competition into the UG (Fischbacher et al., 2002; Roth et al., 1991). Suppose that instead of one there are two responders who simultaneously decide whether to accept or reject a given offer by the proposer. If both responders reject, all three players receive a zero payoff. If both accept, each responder has a 50 percent chance of receiving the proposed amount. If only one responder accepts, the accepting responder receives the proposed amount, the proposer receives the rest and the rejecting responder has a zero payoff. The crucial element in this game is that if one of the responders is willing to accept an offer, the other responder can no longer punish the proposer by rejecting the offer. The impossibility of punishment is like an infinite cost of punishment for the responder. Therefore, one should expect that the responders reject the same offers much less in the competitive situation in comparison with the bilateral UG. Moreover, the reduction of the rejection rate should occur only in those cases in which a responder believes that the competitor accepts the offer. Both facts are fully born out by the data (Fischbacher et al., 2002): In the competitive situation, the rejection

rate is dramatically lower and the whole reduction can be explained by the responders' pessimistic beliefs about the competing responders' behaviour. The evidence also shows that the proposers take advantage of the lower rejection rate and propose much lower offers in the competitive situation. This suggests that the proposers anticipate the lower rejection rate by the responders.

Anecdotal and ethnographic evidence suggests that ethnicity and other group boundaries may exert a strong influence on human altruism. There is a large experimental literature suggesting that even members of "minimal" groups give more favourable treatment to the other members of their group and less favourable treatment to out-group members (Tajfel, 1982; Tajfel et al., 1971). Members of a minimal group share a trivial, purely nominal, social category, e.g. if subjects are allocated to groups according to whether they prefer Kandinski paintings over Klee paintings or vice versa. In a typical minimal group experiment, the members of a group make a series of decisions that determine the payoff between two other subjects – one of these subjects is an in-group member the other one an out-group member. For instance, subjects might have to decide between (5 for subject A, 2 for subject B) and (4 for A, 4 for B). The evidence from such experiments shows that a non-negligible fraction of subjects allocate higher payoffs to in-group members. While this evidence is not contested, the interpretation of the facts is highly controversial. The social identity interpretation (Tajfel, 1982) assumes a universal human motivation for maintaining a positive self-identity. Because self-identity partly derives from social identity, i.e., from identity of the groups to which one belongs, subjects are assumed to be motivated to treat members of their own group in a positively distinct way.

This interpretation has, however, been convincingly challenged by experiments showing that in-group favouritism in minimal groups is reversed if in-group members know that their own payoff is not determined by other in-group members' decisions but exclusively by the decisions of out-group members (Rabbie et al., 1989; Yamagishi et al., 1999). If in-group subjects know that their payoff is determined by other in-group members, a significant number of the in-group subjects behave more altruistically towards other in-group members. If, instead, in-group subjects know that their payoff is determined by out-group members, a significant number of in-group members behaves more altruistically towards the out-group members.

It is important to stress that subjects in these minimal group experiments do not interact in pairs so that direct reciprocation with other in- or out-group members is completely ruled out. Nevertheless, subjects allocate more money to those groups whose members can affect their payoff, suggesting that they expect a kind of generalized reciprocation at the group level. Similar results have been confirmed by prisoners' dilemma games taking place within and across minimal groups (Yamagishi et al., 1999; Yamagishi and Kiyonari, 2000). Subjects typically expect more cooperation from in-group members, which induces them to cooperate more with in-group members. In fact, if one controls for subjects' expectations, the higher level of cooperation when paired with other in-group members vanishes. Thus, in-group favouritism in prisoners' dilemmas has noth-

ing to do with social identity per se, but can be fully explained by the higher expected reciprocation from other in-group members.

Although subjects in minimal groups do not discriminate between in- and out-group members, if one controls for their expectations of reciprocation, it seems plausible that the propensity for altruistic rewarding and punishment in naturally existing groups towards out-group members is lower. Until now, however, there has been no rigorous evidence supporting this conjecture. In a sequential social dilemma conducted in Israel, involving Ashkenazic Jews (European and American immigrants) and Eastern Jews (Asian and African immigrants), both Ashkenazic and Eastern Jews who were in the role of a first-mover cooperated less when they knew that the second-mover was an Eastern Jew (Fershtman and Gneezy, 2001). However, in a dictator game both groups gave the same amount to Ashkenazic and to Eastern Jews, indicating that the differential treatment in the social dilemma is driven by a lack of trust in Eastern Jews. Moreover, if one controls for first-mover cooperation, the Ashkenazic and Eastern second-movers exhibited the same cooperation in the social dilemma. Thus again, if one controls for subjects' expectations of reciprocation (or, in the case of second-movers, for first-mover behaviour) group boundaries have no effect on altruistic behaviour.

2. Proximate theories

Altruistic rewarding and punishment imply that individuals have motives beyond their economic self-interest – their subjective evaluations of economic payoffs differ from the economic payoffs (Thibaut and Kelley, 1959). Take for example the PD. According to the economic payoffs, it is in the self-interest of each player to defect regardless of what the other player does (Figure 4a). In fact, however, many subjects behave as if they prefer the mutual cooperation outcome over the outcome in which they defect and the other player cooperates. Thus, from the viewpoint of individuals with such preferences, the payoff matrix in Figure 4b captures the strategic situation much better. The game in Figure 4b is no longer a PD but a coordination game in which both mutual defection as well as mutual cooperation are in equilibrium. The crucial point is that a strong reciprocator is willing to cooperate if he believes that the opponent cooperates and, therefore, mutual cooperation is in equilibrium if two strong reciprocators are playing the PD. However, since mutual defection is also in equilibrium, it depends on the individuals' beliefs about the other players' actions as to whether the mutual cooperation or the mutual defection equilibrium is played.

Recent results on the neurobiology of cooperation in the PD support the view that individuals experience particular subjective rewards from mutual cooperation (Rillling et al., 2002). If subjects achieve the mutual cooperation outcome with another human subject, the brain's reward circuit (components of the mesolimbic dopamine system including the striatum and the orbitofrontal cortex) is activated relative to a situation in which subjects achieve mutual cooperation with a programmed computer. Moreover, there is also evidence indicating a negative response of the dopamine system if a subject cooperates but

		Player 2	
		Cooperate (C)	Defect(D)
Player 1	Cooperate (C)	20, 20	0, 30
	Defect (D)	30, 0	10, 10

(a)

		Player 2	
		Cooperate (C)	Defect(D)
Player 1	Cooperate (C)	20, 20	-15, 15
	Defect (D)	15, -15	10, 10

(b)

		Player 2	
		Cooperate (C)	Defect(D)
Player 1	Cooperate (C)	20, 20	-30α, 30-30α
	Defect (D)	30-30α, -30α	10, 10

(c)

Figure 4: Subjective evaluation of economic payoffs in the prisoners' dilemma. The first number in each cell denotes the payoff of player 1, the second number is the payoff of player 2. Each player can cooperate (C) or defect (D). **a**. Economic payoffs in the prisoners' dilemma. For any given action of the other player, the economic payoff for each player is higher when he defects. **b**. Example of subjective payoffs in the prisoners' dilemma if both players are strong reciprocators. The subjective payoff from CC is 20 whereas the subjective evaluation of DC is only 15, inducing a strong reciprocator to cooperate if he believes that the opponent cooperates as well. If a player believes that the opponent defects for sure his subjective expected payoff from cooperation is -15 (indicating that a strong reciprocator dislikes being the sucker) whereas his subjective expected payoff from defection is 10 inducing the player to defect as well. Thus, both mutual cooperation (CC) and mutual defection (DD) is an equilibrium. **c**. Subjective payoffs in the prisoners' dilemma if inequity aversion is the motive behind strongly reciprocal behaviour. $\beta > 0$ measures how much a player dislikes advantageous inequity (guilt) whereas $\alpha > 0$ measures how much a player dislikes disadvantageous inequity (envy). If 30 - 30β is smaller than 20 a player prefers CC over DC rendering CC an equilibrium outcome. This holds if β exceeds 1/3. Since $\alpha > 0$, an inequity averse player always prefers to defect if the other player defects rendering DD also an equilibrium outcome. Note that if both players exhibit $\alpha = \beta = \frac{1}{2}$ the subjective payoffs in part (c) are identical to those in part (b) of the figure.

the opponent defects. In another paper (Singer et al., 2004) it is shown that just seeing the faces of people who previously reciprocated cooperation in a sequential social dilemma game activates important parts of the brain's reward circuits. Moreover, not only mutual cooperation seems to involve non-pecuniary rewards but altruistic punishment also seems to be driven by activations in the human reward system. In a recent paper (de Quervain et al., 2004) it has been shown that the nucleus caudate is activated if subjects have the chance to punish free riders in a social dilemma game. The nucleus caudate is a key reward area that is typically activated when rewards are acquired through decisions or actions.
There is a extensive literature in social psychology and economics proposing different social motives behind human altruism (Andreoni, 1990; Frohlich and Oppenheimer, 1984; Loewenstein et al., 1989; MacCrimmon and Messick, 1976; Margolis, 1982; Messick and Sentis, 1985; Thibaut and Kelley, 1959). One potential problem with these approaches is that ex-post, once the result of a particular experiment is observed, it is always easy to "invent" a particular motive or set of motives to "explain" the result. This criticism can be met if it is possible to explain subjects' behaviour in many different situations with a single additional motive or a very small set of additional motives. For example, is it possible to explain that the same individuals often both reward and punish other people. Which motives can explain both the low levels of cooperation in the public good game without punishment and the high cooperation levels in the game with a punishment opportunity? Is it possible to explain using the same motives that responders in the UG reject an offer of, say, 30% of the available money whereas when there is competition they accept offers of 10% or less? Why are responders in the UG tougher bargainers when there are reputation formation opportunities and why do subjects cooperate more in the repeated PD than in the one-shot PD? Fortunately, recent game theoretic models, which combine the existence of social motives with a rational choice approach, indicate that all these phenomena can be parsimoniously explained without inventing new motives for every situation (Bolton and Ockenfels, 2000; Falk and Fischbacher, in press; Fehr and Schmidt, 1999; Levine, 1998).

2.1 Inequity aversion

Theories of inequity aversion (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999; Loewenstein et al., 1989; Messick and Sentis, 1985) assume that a non-negligible percentage of people are both motivated by their own economic payoff **and** dislike outcomes that are perceived as inequitable. The equitability of an outcome is determined by comparisons with reference groups and reference outcomes, which are themselves determined by complicated social comparison processes (Festinger, 1954; Runciman, 1966). Recent theories of inequity aversion (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) assume that if individuals are ahead of the reference outcome they are motivated to help others by transferring resources to those with lower economic payoffs. Yet, if they are behind the reference outcome they are motivated to lower the economic payoff of those ahead (Figure 5). In laboratory experiments, though less often in real life

situations, it is frequently plausible to assume that equality in economic payoffs is the relevant reference outcome. The reason is that experimental subjects enter the laboratory as equals, they don't know anything about each other, and they are allocated to the different roles in the experiment at random. It then seems natural to assume that the reference group is simply the set of subjects playing against each other and the reference outcome is given by the egalitarian outcome, so that inequity aversion boils down to inequality aversion (Figure 5 and Box 1). Inequality averse individuals are willing to incur cost to narrow the difference in economic payoffs between themselves and their relevant reference agents.

Inequity aversion can be thought of as capturing the emotions of guilt and envy. If subjects are better off than the reference outcome they feel guilt, whereas if they are worse off they feel envy. Guilt motivates them to help those who are worse off while envy motivates them to hurt those who are better off. In this view, altruistic rewarding is driven by guilt while altruistic punishment is motivated by envy. Inequality aversion can be parsimoniously modelled by an envy parameter α and a guilt parameter β (Box 1). The larger α, the more an individual suffers from being behind; the larger β, the more the individual suffers from guilt if ahead. The assumption that there are individuals with inequality averse preferences, who rationally chose their actions to meet their goals, can explain why we observe cooperation in one-shot PD's, why responders reject low offers in the bilateral UG, why responders seem to care much less about low offers when competition prevails, and why cooperation often is rather low in the absence of punishment opportunities in public good experiments (Box 1).

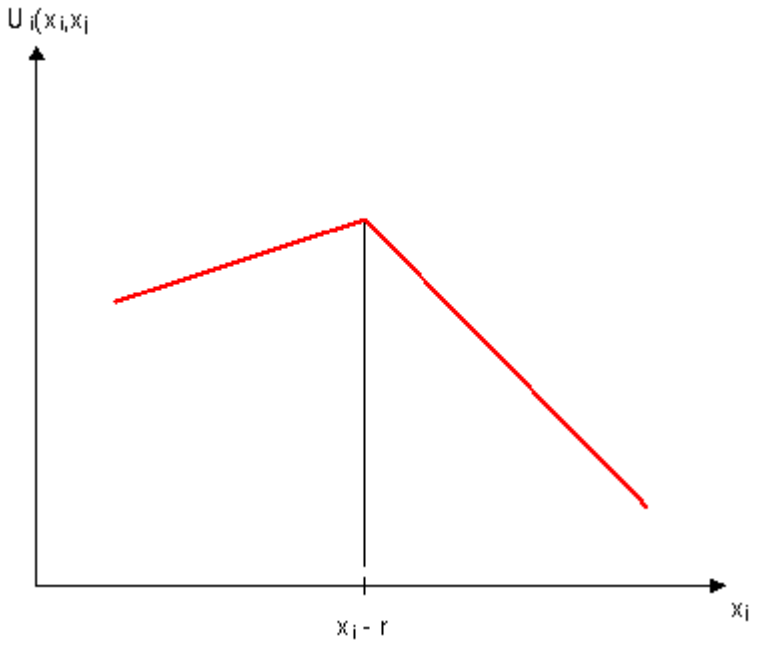


Figure 5: The subjective evaluation (utility) of an inequity averse player i as a function of the economic payoff of reference player j. The economic payoff of player i is fixed at x_i, the economic payoff of player j, x_j, varies along the horizontal axis and r_{ij} denotes the reference outcome for player i relative to player j. If $x_i - r_{ij} > x_j$ an increase in x_j increases i's subjective evaluation of the outcome, i.e., i is willing to transfer resources to j. If $x_i - r_{ij} < x_j$ an increase in x_j decreases i's subjective evaluation of the outcome, i.e., i is willing to reduce the economic payoff of j. The utility of player i is highest if $x_i - r_{ij} = x_j$. For $r_{ij} = 0$ inequity aversion is identical to inequality aversion. The assumption of $r_{ij} = 0$ often makes sense in laboratory experiments. Outside the laboratory, inequity often may not coincide with inequality. Motives depending on non-egalitarian reference outcomes also trigger strongly reciprocal behaviours but the domain in which helping and punishing behaviours occur depends on the reference outcome.

Box 1: Inequity Aversion and Cooperation

The goal of an inequity averse individual i can be formalized as follows. Let x = (x_1, x_2, ..., x_n) denote the economic payoffs of all n players while r_{ij} denotes a reference point of player i relative to player j. Then the objective function of an inequity averse player i is given by $U_i = x_i - \Sigma_{j \neq i} v_i(x_i - r_{ij} - x_j)$. v_i measures player i's disutility from inequity as a convex function of $x_i - r_{ij} - x_j$. v_i equals zero if equity, which is defined as $x_i - r_{ij} = x_j$, prevails. For $r_{ij} = 0$ inequity aversion is tantamount to inequality aversion. Positive and negative deviations from equity cause disutility so that v_i is positive for $x_i - r_{ij} \neq x_j$. If $x_i - r_{ij} > x_j$, v_i is decreasing in x_i and increasing in x_j. If $x_i - r_{ij} < x_j$, v_i is increasing in x_i and decreasing in x_j. Linear, parametric, versions of inequity aversion with $r_{ij} = 0$ have been proposed to enhance mathematical tractability. A version (Fehr and Schmidt, 1999) which explains a wide variety of facts is $U_i = x_i - [\alpha_i/(n-1)] \max \Sigma_{j \neq i}\{x_j - x_i, 0\} - [\beta_i/(n-1)] \max \Sigma_{j \neq i}\{x_i - x_j, 0\}$ with the additional assumptions of with $\beta_i \leq \alpha_i$ and $\beta_i \leq 1$. Here α_i measures the disutility from disadvantageous inequality ($x_j - x_i > 0$) while β_i measures the disutility from advantageous inequality ($x_i - x_j > 0$).

This parametric version of inequity aversion implies that a small minority of selfish players (with $\alpha_i = \beta_i = 0$) can induce a large majority of inequity averse players to contribute nothing to a public good if targeted punishment is not available. Consider a public good situation in which each player has a resource endowment y which can be used for private purposes or spent on a public good G. Denote the contribution of i to the public good by c_i and assume that G is given by the sum of all contributions Σc_i. Let the economic payoff of subject i be given by $x_i = y - c_i + b\Sigma c_i$ where b, which obeys $1/n < b < 1$, measures the pecuniary benefit of the public good. Then, the theorem says that if $b + \beta_i < 1$ player i never contributes anything to G. Moreover, if there are n^o players in the group who obey $b + \beta_i < 1$ and if $n^o > b(n-1)/2$ the unique equilibrium is given by $c_i = 0$ for all group members. Suppose, for instance, that n = 100 and b = 0.1. Then, if all members contribute nothing, each one earns y whereas if all members contribute everything each members earns bny = (0.1)100y = 10y. Despite the ten-fold income gains from full contributions, a minority of 5 selfish players suffices in attaining the unique equilibrium of zero contributions by everybody because the critical threshold b(n-1)/2 is given by (0.1)99/2 = 4.45. Thus, even if 95 players are very inequity averse, the unique equilibrium is given by zero contributions.

The situation fundamentally changes if direct punishment is possible. Suppose that player i can reduce the income of j by imposing punishment p_{ij} on j and that i has to pay kp_{ij} ($k > 0$) for this. Then it can be shown (Fehr and Schmidt, 1999) that full contributions by everybody is an equilibrium if there is a subgroup of n' "conditionally cooperative punishers" and if the cost k for the punishers are not too large. The conditionally cooperative punishers are conditional cooperators (i.e. their preferences obey $b + \beta_i > 1$) who punish the defectors. The cost k have to obey the condition $k < \alpha_i/[(n-1)(1+\alpha_i) - (n'-1)(\alpha_i+\beta_i)]$ for all n' conditionally cooperative punishers.

A subject who feels sufficient guilt in case of unilateral defection in the PD (high β-value) does not like to defect if the other player cooperates. Inequality averse subjects are, therefore, conditional cooperators – they cooperate as long as they believe that the opponent also cooperates. Hence, if two inequality averse players are matched, mutual cooperation is an equilibrium in the PD (Figure 4c). Moreover, if a responder in the UG dislikes earning less than the proposer, the responder prefers to reject low offers. Likewise, an inequity averse third party in the TPG punishes a greedy player A who gives nothing to the poor player B. Since inequality averse subjects value both their own payoff **and** equality, this approach can also explain why subjects in the PD are more likely to cooperate if the probability of repeated interactions increases. Likewise, the theory can explain why responders in the UG reject more often if they can gain an individual reputation. The reason is simply that without reputation, only the equity motive triggers rejections whereas if reputation formation is possible, both the equity motive and self-interest triggers rejections. Yet, how can inequality aversion explain that subjects reject much less under responder competition? The key factor here is that once responder A believes that the competing responder B accepts a low offer, responder A can no longer hurt the proposer by rejecting the offer. This is so because if B accepts, the proposer's proposal will be implemented. Therefore, even for an inequality averse responder with a very high envy parameter α it does not make sense to reject a low offer if he believes that the competing responder accepts. Inequality aversion also explains why those who cooperate in public good games punish the defectors: Through punishment, the cooperators can remove the difference in economic payoffs between themselves and the defectors. To explain the breakdown of cooperation in the n-player public goods game without punishment opportunities, the heterogeneity in players' preferences is crucial: The selfish players will always defect in the absence of punishment and, therefore, inequality averse players, who condition their cooperation on the other players cooperation will also defect (Box 1).

2.2 Reciprocal fairness

Although inequity aversion motives can explain a large and diverse set of facts, the concept has also its limits. One important limitation arises if the punishment cost for the punisher and the punished subject are the same. In this case, the punisher cannot change the payoff difference between himself and the punished person. Therefore, an inequity averse subject should never punish in this situation. There is, evidence, however, that between 20 and 30 percent of the subjects punish even if they cannot change the payoff difference (Falk et al., in press). A further limitation of inequity aversion can be illustrated by the following two conditions in a simplified UG (Brandts and Sola, 2001; Falk et al., 2003) where 10 MUs have to be divided. In one condition the proposer can only propose either (5.5) or (8.2). In the other condition he can only propose either (10.0) or (8.2). An inequity averse responder will reject (8,2) regardless of whether (5.5) or (10.0) was the alternative to (8.2). However, an offer of (8.2) is likely to be interpreted by the responders as more unfair if (5.5) instead of (10.0) was the

alternative. In fact (Falk et al., 2003), the responders reject the (8.2) proposal with probability.09 if (10.0) is the alternative proposal, whereas if (5.5) is the alternative the (8.2) offer is rejected with probability.44. This suggests that the perceived fairness of an outcome depends not only on the outcome itself but also on the available set of alternative outcomes that could have been chosen by the proposer. A plausible way to interpret this is to say that the proposer's intentions are important for the responder's fairness judgement. If the proposer could have chosen the egalitarian distribution (5.5), the offer (8.2) signals greater unfair intentions than if the alternative was the (10.0) offer. Another interpretation is to say that a proposer who chooses (8.2), although (5.5) was available, reveals that he is a greedier person.

These two interpretations provide the intuition for two formal theories of reciprocal fairness (Box 2). One theory is based on the view that strong reciprocators' goals are to reward fair and to punish unfair individuals (Levine, 1998). The other theory (Dufwenberg and Kirchsteiger, 2004; Rabin, 1993) rests on the motivational assumption that strong reciprocators want to reward fair and to punish unfair intentions and that intentions can be assessed by evaluating the actual choice in the light of possible alternatives. The difference is that intentions depend on the specific situation – a fair person may sometimes also exhibit unfair intentions – whereas in the first approach, fairness is viewed as a personality characteristic.

Box 2: Reciprocal Fairness and Cooperation

All models of reciprocal fairness have the feature that the economic payoff of other players is weighted positively if they are considered to be kind and negatively if they are considered unkind. Formally, the different models can be represented by the utility function $U_i = x_i + \sum_j v_i(\kappa_i^j) \cdot x_j$ where U_i is the utility of player i and x_i is the material payoff of player i. The term κ_i^j measures player j's kindness towards player i. Player i evaluates κ_i^j with a monotonous valuation function v_i which can be positive or negative. If v_i is positive, player j's payoff is valued positively which may lead to altruistic rewarding. If v_i is negative, j's payoff is valued negatively which may lead to altruistic punishment. For selfish players v_i is always zero. The different models differ in their definitions of kindness.

Personality based reciprocal fairness (Levine, 1998). Assume that players differ in how altruistic they are. Their degree of altruism can be captured by the parameter α_i. Personality based theories assume that people predict other individuals' altruism parameter. They respond by altruistic rewarding or altruistic punishment depending on their prediction of others' altruism parameters. More formally, the utility payoff of such players is given by

$$U_i = x_i + \sum_j \frac{\alpha_i + \lambda_i \cdot \alpha_j}{1 + \lambda_i} x_j$$

where $\alpha_i \in (-1, 1)$ captures player i's altruistic motivation and the reciprocity parameter $\lambda_i \in [0, 1]$ measures player i's preference for reciprocation. Here kindness κ_i^j is defined by player j's altruism parameter α_j. The valuation function v_i is given by $v_i(\kappa_i^j) = (\alpha_i + \lambda_i \kappa_i^j)/(1 + \lambda_i) = (\alpha_i + \lambda_i \alpha_j)/(1 + \lambda_i)$. This model has two key properties: First, the higher α_i, the more player i values the other players' payoff. If $\alpha_i < 0$, player i is even spiteful, i.e., he prefers to reduce the other player's economic payoff. Second, the higher the altruism parameter of the other player j, the more a reciprocal player i (with $\lambda_i > 0$), values player j's economic payoff. In a public goods game, players with a sufficiently high α_i will cooperate. Furthermore, everything else equal (reciprocity parameter λ_i and belief about the other players), players with the α_i cooperate at a higher level. Therefore, players who have contributed little or nothing will be identified as less altruistic. Sufficiently reciprocal players weigh the payoff of these players negatively and, if punishment is possible, they punish the defectors.

Menu based reciprocal fairness (Rabin, 1993). In menu based models, j's kindness is determined by the actual choice of j in comparison to the alternatives (the available menus). Let A_i^j denote the set of available alternatives, which determine the possible payoffs player i can get depending on player j's choice. Let π_i^L be the lower payoff limit of A_i^j and π_i^H the upper limit of A_i^j. We define the fair payoff as $\pi_i^F = \left(\pi_i^H + \pi_i^L\right)/2$. Let π_i^A be the payoff of player i given the actual choice of player j. The kindness κ_i^j of player j toward i is defined as 0 if $\pi_i^H = \pi_i^L$ and as $2\left(\pi_i^A - \pi_i^F\right) / \left(\pi_i^H - \pi_i^L\right)$ otherwise. This expression

is always between –1 and +1. The evaluation function in this model is simply the multiplication of κ_i^j with an individual reciprocity parameter $\rho_i \geq 0$, which measures the weight of the reciprocity motive. Therefore, in the two-player case, the utility of player i is defined as $U_i = x_i + \rho_i \kappa_i^j x_j$ which is determined by the actions and the beliefs of the players. A reciprocity equilibrium is then defined as a combination of actions and beliefs in which first, all players choose a strategy to maximize their utility and second, beliefs match the actual behaviour.

In a PD, cooperation has a kindness value of +1 and defection a value of -1 (independent of the beliefs). Therefore, if the opponent cooperates and if ρ_i is sufficiently large, player i also cooperates so that there is an equilibrium with mutual cooperation. In contrast, if the opponent defects, his kindness equals -1 and therefore a reciprocal player also defects. If there is a punishment stage, sufficiently reciprocal players punish defectors because defectors' kindness is negative so that their payoff is weighted negatively.

Outcome based reciprocal fairness (Fehr and Schmidt, 1999). Outcome oriented models mimic reciprocal fairness by inequity aversion. Kindness corresponds to $\kappa_i^j = x_i - x_j$ and the evaluation function is given by

$$v_i(\kappa_i^j) = \begin{cases} +\beta_i/(n-1) & if \quad \kappa_i^j > 0 \\ 0 & if \quad \kappa_i^j = 0 \\ +\alpha_i/(n-1) & if \quad \kappa_i^j < 0 \end{cases}$$

Based on this definition, outcome based reciprocal fairness can be transformed into inequity aversion by assuming a utility function $U_i = x_i + \sum_j v_i(\kappa_i^j) \cdot (x_j - x_i)$. Note, however, that kindness now does not weight the other player's payoff but weights inequity. This is the reason why inequity aversion predicts no punishment if the cost of punishment is identical for the punisher and the punished individual.

Both the intention-based and the personality-based approaches are based on powerful intuitions, which can rationalize why people reject low offers in the UG and why they are conditionally cooperative in the PD. It also seems possible that the reciprocal fairness approach is capable of explaining why rejections are much less frequent in the UG with responder competition, why responders reject more often if they can gain an individual reputation for being tough, and why cooperation in the PD increases if the probability of repeated encounters is higher. However, formally modelling intention-based fairness is a tricky business, and up to now no model which is mathematically easily tractable exists which would permit precise and empirically correct predictions across many different experiments. In addition, there is evidence that it is not the unfairness of the intentions alone which drives altruistic rewarding and altruistic punishment (Blount, 1995; Falk et al., 2003). If, e.g. the proposer in the constrained ultimatum game is forced to choose the alternative (8.2), because no other alternative is available, still 18 percent of the responders reject the (8.2) outcome. The limits of both the intention-based reciprocity approach and the inequity aversion approach have motivated models combining the two concepts (Falk and Fischbacher, in press). The combined approach can explain what can be clarified by inequity aversion and by intention-based reciprocal fairness and it overcomes the above-mentioned limits of both approaches.

The personality-based approach towards reciprocal fairness (Levine, 1998) has proven to be more amenable to tractable formalization (Box 2). This approach also provides a natural explanation for the existence of third party punishment and third party rewarding: third parties punish those who are greedy and they reward those who have helped others in the past. In contrast, the currently available intention-based approaches (Dufwenberg and Kirchsteiger, 2004; Rabin, 1993) have difficulties in explaining third party rewarding and punishment because they are built on the reciprocal fairness motives between two interacting parties. Yet, by definition, third parties did not have previous interactions with the rewarded or punished persons.

3. Evolutionary Origins

What are the ultimate origins behind the rich patterns of human altruism described above? And, equally important, to what extent are the different evolutionary forces behind human altruism capable of explaining cooperation in relatively large groups? Before we go into the details, it is important to stress a basic truth: any model which successfully explains the evolution of altruistic behaviours must deliver a mechanism that ensures that the benefits of altruistic acts fall predominantly on other altruists. If such a mechanism is absent, altruism cannot evolve.

3.1 Reciprocal altruism

Reciprocal altruism in the form of tit-for-tat or similar cooperation-sustaining strategies in the repeated PD provides such a mechanism. Here cooperation

can evolve because only the two players involved in the PD benefit from mutual cooperation. The experimental evidence also unambiguously shows that in two person interactions, subjects cooperate more if future interactions are more likely (Andreoni and Miller, 1993; DalBo, 2003; Gächter and Falk, 2002). Although the evolutionary paradigm of reciprocal altruism rests on powerful intuitions and explains cooperation in small kin-based groups rather well, the paradigm still suffers from several limitations: first, the interacting individuals are typically forced to stay together for a random number of periods (Hammerstein, 2003). This assumption is not only violated by many, if not most, animal interactions but it is also clearly violated in the case of genetically unrelated humans. Throughout evolutionary history, humans have almost always had the option of breaking off interactions with unrelated individuals. This does not mean that there may have been considerable obstacles in leaving a relationship; yet, unless the available outside options and individuals' decisions to stay in or to leave a relationship are modelled explicitly, it is impossible to study their impact on the evolution of altruistic behaviour. Second, with a few exceptions (Bendor and Swistak, 2001; Boyd and Richerson, 1988, 1992) the evolutionary analysis of repeated encounters has been largely restricted to two-person interactions. While this may perhaps be justified for non-human animals, the human case clearly demands the analysis of larger groups. Unfortunately, the evolutionary success of tit-for-tat like strategies of conditional cooperation is extremely limited even in relatively small groups exceeding two persons. It turns out that in n-person PDs the only conditionally cooperative, evolutionarily stable strategy prescribes cooperation only if all other group members cooperated in the previous period. Moreover, the basis of attraction of this strategy is extremely small, rendering the achievement of cooperation in larger groups very unlikely (Boyd and Richerson, 1988). We simulated the evolution of cooperation in multi-person PDs when group members repeatedly interact with each other over many periods (Figure 6). It turns out that substantial rates of cooperation can only be maintained in two-person groups and, regardless of the initial conditions, cooperation quickly drops to zero in groups with more than two members. The reason for this result is that a few selfish players in a group suffice to undermine the cooperation of the conditional cooperators. Thus, repeated interactions plus the existence of strategies which condition cooperative behaviour on past outcomes (i.e., reciprocal altruism) are unlikely to be an evolutionary explanation for human cooperation in larger groups.

Third, reciprocal altruism cannot provide adaptive explanations of altruistic rewarding and punishment. It is, however, quite common to argue that strong reciprocity is a maladaptive consequence of reciprocal altruism (Binmore, 1998; Johnson et al., 2003). In this view, subjects apply behavioural rules like tit-for-tat, which might make sense in repeated two-person interactions, maladaptively to one-shot encounters. In the evolutionary past, so the argument goes, one-shot interactions were rare and therefore subjects' rules are not fine-tuned to the one-shot situation. This argument faces many problems. It is contradicted by evidence from many ethnographic accounts of hunter-gatherer societies, which indicate that fitness relevant interactions with strangers have been relatively

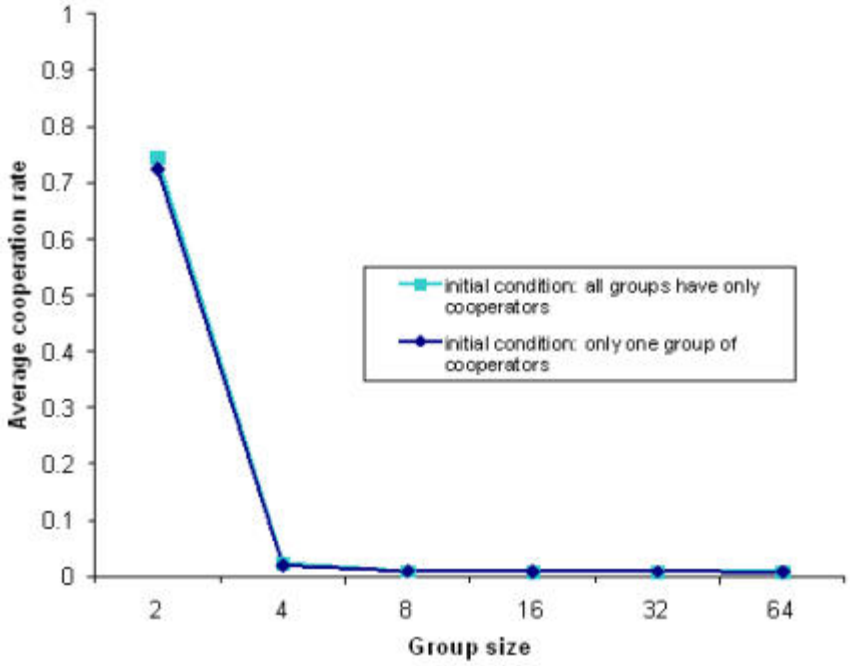


Figure 6: Cooperation under reciprocal altruism in a multi-person Prisoners' Dilemma. We conducted simulations of the evolution of cooperation in 64 groups of fixed size n with n ranging from 2 to 256. The figure shows, for each group size, the average cooperation rate in 100 independent simulations over the last 1000 of 2000 generations. Cooperation has a fitness cost of c = 0.2 for the cooperator but each of the n-1 other players in the group reaps a fitness benefit of 1/(n-1) from a cooperative act. There are full defectors and conditional cooperators (i.e., reciprocal altruists) of varying degrees x. A reciprocal altruist cooperates in period t if at least x percent of the other group members cooperated in t-1. The figure shows that under reciprocal altruism cooperation can only be maintained in two-person groups. This holds regardless of whether we start with all groups being fully cooperative or only one group being fully cooperative. Group members interact repeatedly, i.e., the probability of a further period is 0.9. Strategies are implemented with an error probability of 2 percent. When the groups (a generation) cease to exist (with probability 0.1) the fitness of each individual j is compared to a randomly chosen individual k from the overall population and the probability that strategy j is replicated is given by (f_j/f_j+f_k). We chose this selection mechanism to render the results in Figure 6 comparable to the results in Figures 7a and 7b. If, for selection purposes, we only compare the relative fitness of randomly chosen individuals from the same group, or if we allow for probabilistic relative fitness comparisons between in-group and out-group member, cooperators do even worse. After selection, new groups are randomly formed.

frequent (Fehr and Henrich, 2003). The argument is also contradicted by the experimental facts: subjects cooperate significantly more if they know that they will interact with the same individual again in future periods (Andreoni and Miller, 1993; DalBo, 2003; Gächter and Falk, 2002). Moreover, the crucial point is not that defection is the best strategy in one-shot encounters, while cooperation always pays in repeated encounters. If the present gains from defection are large enough, the best strategy for a selfish individual is to defect even if the probability of a repeated encounter is as high as or higher than 0.9. Plausibility and the ethnographic evidence suggest that – depending on the interaction partner – humans faced many different probabilities of repeated encounters so that often situations arose in which defection was the best strategy. Indeed, the very fact that humans seem to have excellent cheater detection abilities (Cosmides and Tooby, 1992) suggests that, despite many repeated interactions, cheating has been a major problem throughout human evolution. Therefore, humans' behavioural rules are likely to be fine-tuned to the variations in cheating opportunities. Finally, the maladaptation argument cannot explain why non-human primates, who live in groups with many repeated interactions, do not exhibit strongly reciprocal behaviours. If humans "mistakenly" cooperate in encounters with a "low" frequency of future interactions why do other primates not make the same "mistake"?

3.2 Reputation-seeking

Evolutionary approaches of reputation-based cooperation represent important steps beyond reciprocal altruism (Alexander, 1987; Gintis et al., 2001; Leimar and Hammerstein, 2001; Nowak and Sigmund, 1998a, b; Zahavi, 1995). In the indirect reciprocity model (Alexander, 1987; Leimar and Hammerstein, 2001; Nowak and Sigmund, 1998a, b), third parties reward individuals with a good reputation, i.e., individuals with a good image score, if they can themselves acquire a good image score by rewarding. The idea behind this approach seems compelling: people help unrelated others with whom they have no further future interactions because by helping they increase the probability that they themselves receive help in the future. Although the general idea of reputation-based cooperation is attractive, there are some unsolved problems in the indirect reciprocity approach that point towards the need for further research. First, the approach produces long-run helping rates of roughly 40 percent if the recipient's benefit is four times the donor's cost, provided that all individuals live in isolated groups without migration. If, however, genetic mixing between the groups occurs, helping rates decline dramatically and approach zero (Leimar and Hammerstein, 2001). It would be an important step forward if the indirect reciprocity approach could be modified in such a way that significant helping rates could be maintained under reasonable assumptions about migration between groups. Second, it is still an open question how best to model the concept of a good reputation. Should an individual who does not help a person with a bad reputation lose her good reputation? Currently the image scoring approach (Nowak and Sigmund, 1998a, b) gives an affirmative answer to this question while oth-

ers take a negative position (Leimar and Hammerstein, 2001). In our view this question is intrinsically related to a society's prevailing norms, which are themselves the product of evolutionary forces. This raises the question whether it is possible to model the evolution of reputation formation in dyadic interactions independently of the norms that prevail in a society. In our view the notion of normatively appropriate behaviour – from which a good reputation derives – is itself subject to the forces of evolution.

Until recently, the indirect reciprocity approach has been limited to dyadic cooperation. Therefore, it could not explain cooperation in larger groups. Yet, recent experiments that connect the public good game with an indirect reciprocity game point towards a potential solution (Milinski et al., 2002b). Individuals typically interact in many different domains, e.g., they contribute to public goods and they have dyadic interactions with others. If the reputation created by higher contributions to a public good also increases the probability of receiving help in dyadic interactions, there is an incentive to contribute to the public good. A recent paper (Panchanathan and Boyd, 2004) indeed showed that this intuition can be made rigorous. Large-scale cooperation is possible if cooperation in the indirect reciprocity game is only given to subjects who also cooperate in the public goods game. As individuals who do not cooperate in the public goods game are not helped in the indirect reciprocity game there is an economic incentive to cooperate in the public goods game.

However, Panchanathan and Boyd (2004) also point out that there are many non-cooperative equilibria in their model. Thus, in the absence of a mechanism that ensures that the population converges to the cooperative equilibria, there is no guarantee that reputation-based cooperation evolves in larger groups. One such mechanism could be competition between groups with different social norms because groups who successfully link the helping decision in the indirect reciprocity game with individuals' behaviour in the public goods game are better able to solve their public goods problems. Group competition therefore does not serve as a mechanism for offsetting within-group selection pressures against cooperative individuals but is merely a device for the selection of cooperation-enhancing social norms.

Reputation-based approaches do not provide adaptive explanations for strong reciprocity. Strong reciprocators reward and punish others even when no reputation whatsoever can be acquired. A common speculation among evolutionary psychologists is that humans have hard-wired mental modules that always induce them to behave as if their reputation is at stake. In this view, strong reciprocity is a maladaptive consequence of reputation-based evolutionary forces because, so the argument goes, in the evolutionary past interactions took place exclusively in small groups where individuals' reputations were always at stake. Apart from the fact that there is no precise model that provides a foundation for this view, the experimental evidence also contradicts the argument. From indirect reciprocity games (Engelmann and Fischbacher, 2002) and our UGs with and without reputation formation, we know that subjects respond quickly to opportunities for reputation formation. They punish and reward significantly more when this allows them to gain a favourable reputation. In addition, the ethno-

graphic evidence from hunter-gatherer societies negates the idea that humans **always** behave as if their reputation were at stake (Fehr and Henrich, 2003). It is obvious that in a coincidental interaction with a stranger, an individual's reputation is much less at stake than in an interaction with kin or close friends. Since the ethnographic evidence indicates that humans found themselves in situations with different reputation formation possibilities, it is reasonable to assume that their behavioural rules are fine-tuned to these differences. Therefore, altruistic rewarding and punishment is unlikely to be a maladaptive consequence of reputation-based evolutionary forces.

Costly signalling theory also provides a reputation-based ultimate explanation for altruistic behaviour (Gintis et al., 2001; Zahavi, 1995). According to this approach, individuals signal favourable, yet unobservable, traits by altruistic acts, which render them preferred mating partners. The assumption behind this theory is that individuals with better traits have smaller signalling costs, i.e., smaller costs of altruistic acts. Thus, those with better traits are more likely to signal, which allows the inference that those who signal have better traits on average. The advantage of this approach is that it could, in principle, explain contributions to n-person public goods. The main weakness is that the signalling of unobservable traits need not occur by altruistic acts but can also take other forms. The approach, therefore, generates multiple equilibria – in some equilibria, signalling occurs via altruistic behaviour, in others signalling does not involve any altruistic acts. Therefore, this theory has difficulties explaining human altruism unless it is supplemented with some other mechanisms. One such mechanism might be cultural group selection (Gintis et al., 2001). If groups in which signalling takes place via altruistic behaviour have better survival prospects, selection between groups favours those groups which have settled at a pro-social within-group equilibrium. In this context it is quite important to note that the traditional arguments against group selection do not apply because altruistic signalling is a within-group equilibrium so that there is no within-group selection against the altruistic signallers. However, so far there is neither experimental evidence suggesting that signalling takes pro-social forms nor a dynamic analysis showing that altruistic signallers can evolve when rare. In addition, costly signalling does not provide an adaptive account of strong reciprocity.

3.3 Altruists with green beards

A potential evolutionary explanation for strong reciprocity is to assume that individuals who reward and punish altruistically have observable characteristics ("green beards") that distinguish them from non-altruists (Frank, 1988; Robson, 1990; see Fehr and Fischbacher, and Frank in this volume). It is then in the self-interest of any individual to cooperate with a green beard because non-cooperation will be punished. Thus, altruistic punishment evolves because the punishers directly benefit from their observable willingness to punish. Even in the absence of punishment opportunities, green beards favour the evolution of cooperation because individuals can condition their cooperation to the existence

of a green beard. If an altruist meets a selfish individual without a green beard he defects, if he meets a green beard he cooperates. In this way, the benefits of altruistic behaviour are only reaped by the altruists themselves so that universal cooperation evolves. The green beard argument can be made more realistic by assuming that it takes some effort or that there is only a positive probability of detecting a strong reciprocator (Frank, 1988). Yet, the crucial assumption behind this approach is that there are no selfish mutants with green beards. As soon as one allows for such mutants, the argument breaks down because the mutants reap the same benefits as the altruists but do not bear the cost of altruistic acts. Therefore, any convincing model justifying the green beard argument must allow for such mutants.

In PDs, experimental subjects are in fact capable of predicting other individuals' cooperation rate better than chance if they are given the opportunity to communicate face-to-face and to make promises before the PD is played (Frank et al., 1993). This has been taken as evidence in favour of the hypothesis that humans are able to distinguish true cooperators from those who only pretend to cooperate. There are, however, two objections which question this argument. First, there is extensive literature showing that researchers are able to detect other people's lies with high probability using scientific methods, but that most humans are unable to detect lies better than chance (Ekman and O'Sullivan, 1991; Frank and Ekman, 1997). The inability of humans to detect lies holds for a wide variety of situations, including those where stakes are high. In the most convincing experiment (Ekman and O'Sullivan, 1991) the researchers showed the participants videos from lying and non-lying subjects. They ensured that it was objectively possible, by applying scientific tools, to predict the liars with high probability by watching the videos. The data indicates that 70 percent of the subjects – including subjects who can be expected to be experts, such as federal polygraphers, robbery investigators, judges and psychiatrists, are unable to detect lies significantly better than chance.

Second, the existence of subjects with a preference for conditional cooperation renders predictable behaviour in the one-shot PD if the players have the opportunity to signal their behavioural intentions in a communication stage. This is shown in more detail in the other paper by Fehr and Fischbacher in this volume. Moreover, in the associated paper we also show that a higher share of players with preferences for conditional cooperation increases the predictability of the players' behaviour. Intuitively, the reason for predictability is that individuals with, say, inequity averse or reciprocally fair preferences not only prefer mutual cooperation over unilateral defection, but they may also prefer mutual defection over the DC-outcome (Figure 4c). Therefore, if there is a chance that the opponent will defect, they also prefer to defect. Yet, since there is also a chance that the opponent might cooperate, they want to avoid the DC-outcome by signalling that they defect. Since it is never in the interest of a subject to signal defection when he does not really want to defect, the defection signal is credible, thus rendering subjects' behaviour predictable. It is important to stress that predictability of behaviour does not mean that the players' preferences are predictable. Neither does predictability of behaviour imply that the

strong reciprocators reap higher economic payoffs than the egoists who counterfactually signal cooperation. In fact, the lying egoists earn the most implying that predictability of behaviour in one-shot PD experiments with communication cannot be taken as evidence for the green beard approach towards the evolution of human cooperation.

3.4 Gene-Culture co-evolution

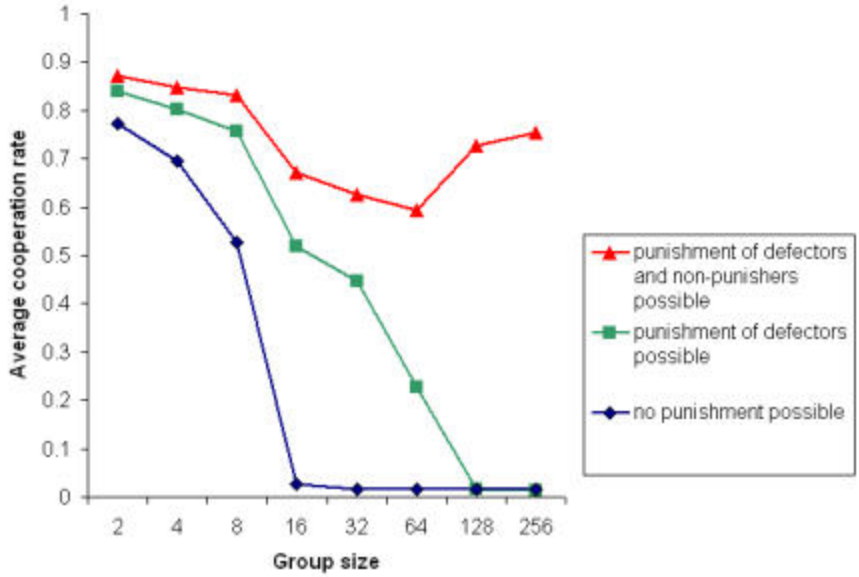
The birth of modern socio-biology is associated with scepticism toward genetic group selection of altruistic traits (Williams, 1966); although possible in theory it has been deemed unlikely to occur empirically. The key argument is that group selection can at best be relevant in small isolated groups because migration in combination with within-group selection against altruists is a much stronger force than selection between groups. The migration of defectors to groups with a comparatively large number of altruists plus defectors' within-group fitness advantage quickly removes the genetic differences between groups so that group selection has little effect on the overall selection of altruistic traits (Aoki, 1982; Rogers, 1990). Consistent with this argument, genetic differences between groups in populations of mobile vertebrates like humans is roughly what one would expect if groups were randomly mixed (Long, 1986). Thus, purely genetic group selection is, like the gene-based approaches of reciprocal altruism and indirect reciprocity, unlikely to provide a satisfactory explanation of strong reciprocity and of large-scale cooperation among humans.

However, the same arguments, which seem plausible when applied to pure genetic group selection, may not be relevant to the selection of culturally transmitted traits or to models of gene-culture co-evolution. Cultural transmission occurs through imitation and teaching, i.e., through social learning. There are apparently large differences in the cultural practices of different groups around the world and ethnographic evidence indicates that even neighbouring groups are often characterized by very different cultures and institutions (Kelly, 1985). Although the existence of culture is perhaps not unique to humans, the degree to which human behaviour is affected by culture is probably unique. In particular, culturally determined behavioural norms and social institutions that are sustained by punishment seem to be uniquely human. Norms and institutions like, e.g., food-sharing norms or monogamy, are important because they weaken the within-group selection against altruistic traits (Bowles et al., 2003). Thus, if one could show that altruistic punishment and rewarding among genetically unrelated individuals can evolve through the evolution of social norms and institutions, one could explain why strong reciprocity seems to be limited to humans.

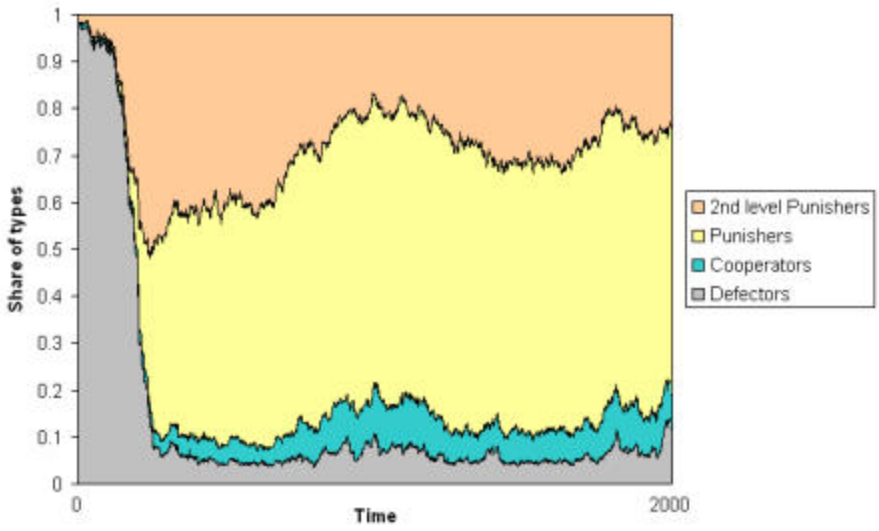
Recent theoretical models of cultural group selection (Boyd et al., 2003; Henrich and Boyd, 2001) or of gene-culture co-evolution (Bowles et al., 2003; Gintis, 2003) could provide a solution to the puzzle of strong reciprocity. The interaction between cultural group selection and altruistic punishment (Boyd et al., 2003; Henrich and Boyd, 2001), for example, could solve two problems at once: The evolution of altruistic punishment and the evolution of cooperation in relatively large groups (Figure 7a and 7b). For the same reasons that speak against purely

genetic group selection, cultural group selection alone is not capable of generating cooperation in large groups (Figure 7a). Yet, when the altruistic punishment of non-cooperators and non-punishers is possible, punishment evolves and cooperation in much larger groups can be maintained. This is for three reasons: first, altruistic punishment of non-cooperators in combination with the imitation of economically successful behaviours prevents the erosion of group differences with regard to the relative frequency of cooperating members. If there are a sufficient number of altruistic punishers, the cooperators do better than the defectors since the latter are punished. Therefore, cooperative behaviour is more likely to be imitated. This fitness disadvantage of defectors also implies that the migration of defectors has much less effect on the evolution of cooperation than in genetic models. Second, the altruistic punishment of non-punishers implies that those who punish non-cooperators do better than pure cooperators who never punish, thus favouring the replication of strategies that punish defectors. Third, when cooperation in a group is wide-spread, altruistic punishers have only a small or no within-group disadvantage relative to pure cooperators who do not punish. In the limit, when everybody cooperates, punishers incur no punishment costs at all and thus have no disadvantage. When no defectors are present, only mutants or errors cause the necessity to punish. Thus, small cultural group selection effects suffice to overcome the small cost disadvantage of altruistic punishers. Moreover, the cost disadvantage of altruistic punishment declines geometrically for higher order punishment (Henrich and Boyd, 2001). Taken together, these forces imply that strategies punishing defectors and non-punishers are maintained at high frequency (Figure 7b).

To what extent is there evidence for the role of culture and group conflict in human altruism? There is strong evidence from intergenerational UGs and trust games that advice by others affects altruistic punishment and altruistic rewarding (Schotter, 2003). In these experiments, successive pairs of players play a trust or an ultimatum game. Subjects who have already played the game can give written advice to the player in the next pair who is in the same role. Advice givers have a direct interest in the behaviour of their successor because they receive half the money earned by the successors. The experiments show that advice is followed and increases altruistic punishment and altruistic rewarding relative to a control condition with no advice. Recent intergenerational public good games with advice giving indicate that later generations achieve significantly higher cooperation levels even in the absence of punishment opportunities (Chaudhuri and Graziano, 2003). UGs and dictator games with children of different ages show that older children are more generous and more willing to punish altruistically (Harbaugh et al., 2000). Although these changes in children's behaviour could be a result of genetic developmental processes, it seems at least as plausible to assume that they are a product of socialisation by parents and peers. Why, after all, do parents invest so much time and energy into the proper socialisation of their children? Perhaps the strongest evidence for the role of cultural norms comes from a series of experiments in 15 small-scale societies (Henrich et al., 2001). There are decisive differences across societies in the behaviour of proposers and responders in the UG. Some tribes like the Hazda from



(a) Figure 7a shows, for each group size, the average cooperation rate in 100 independent simulations over the last 1000 of 2000 generations. An individual who punishes incurs cost of k = 0.2 and the total cost for all punished individuals is p = 0.4. Figure 7a indicates that in the absence of altruistic punishment, no cooperation can be maintained for groups of size 16 or larger. If the altruistic punishment of non-cooperators is possible, substantial cooperation rates can be maintained up to groups of size 32. If, in addition, the altruistic punishment of non-punishers is possible, very high cooperation rates can be maintained for groups up to 256 members. The model underlying Figure 7a differs from the model underlying Figure 6 as follows: first, there are no repeated interactions, ruling out any form of reciprocal altruism. Second the model allows for the following strategies: defectors, pure cooperators who never punish, cooperators who punish non-cooperators (i.e., 1^{st} level punishers), cooperators who punish non-cooperators and non-punishers (i.e., 2^{nd} level punishers). Third, all graphs are based on the assumption that, initially, there is only one group of 1^{st} level or 2^{nd} level punishers. Fourth, after each period player j imitates the strategy of player k with probability (f_k/f_j+f_k). k is a member of the own group with probability 0.8 and a member of another randomly chosen group with probability 0.2. Fifth, after individual within-group selection took place, groups are randomly matched and a conflict occurs with probability 0.05. In case of a conflict, group j replaces group k with probability $\frac{1}{2} + \frac{1}{2}(C_j - C_k)$ where C_j is the frequency of cooperators in group j. Group replacement mimics the fact that the loosing group takes over the cultural traits and institutions of the winning group.



(b) Figure 7b shows the evolution of strategies in a typical simulation when altruistic punishment of non-cooperators and non-punishers is possible (n = 256). Initial conditions ensure that in 63 groups only defectors exist whereas in one group there are only altruistic 2^{nd} level punishers. Already after roughly 200 periods the share of defectors is below 10 percent. The share of pure cooperators varies between 5 and 10 percent whereas the share of 1^{st} level punishers stabilizes after 1000 periods at roughly 50 to 60 percent. After period 1000, the share of the 2^{nd} level punishers varies between 20 and 30 percent.

Figure 7: Cooperation under altruistic punishment and group conflicts in a multi-person Prisoners' Dilemma. As in Figure 6 we simulate the evolution of cooperation in 64 groups of fixed size n with n ranging from 2 to 256. Cooperation has a fitness cost of c = 0.2 for the cooperator but each of the n-1 other players in the group reaps a fitness benefit of 1/(n-1) from a cooperative act.

Tanzania exhibit a considerable amount of altruistic punishment whereas the Machiguenga from Peru show little concerns about fair sharing. Thus, taken together, there is fairly convincing evidence that cultural forces exert a significant impact on human altruism.

Yet, what is the evidence for cultural *group selection*? There is quite strong evidence that group conflict and warfare were wide-spread in foraging societies (Jorgensen, 1980; Otterbein, 1985). There are also examples (Kelly, 1985; Soltis et al., 1995) suggesting that group conflict contributes to the cultural extinction of groups because the winning groups force their cultural norms and institutions on the losing groups. However, although these examples are suggestive, they are not conclusive so that further evidence has high value.

If cultural group selection was a significant force in evolution, then the human propensity to reward and punish altruistically should be systematically affected by inter-group conflicts. In particular, people should be more willing to punish defectors if defection occurs in the context of a group conflict. Likewise altruistic cooperation should be more prevalent if cooperative acts contribute to success in a group conflict. There is evidence from inter-group conflict games indicating that altruistic cooperation in PDs indeed increases if the PD is embedded in an inter-group conflict (Bornstein and Ben-Yossef, 1994). However, in the absence of punishment opportunities the existence of inter-group conflicts does not prevent the decay of cooperation over time (Bornstein et al., 1994) and so far there is no evidence showing that inter-group conflicts increase altruistic punishment. There is also no evidence for the punishment of non-punishers in public good games. In view of the predicted strong impact of this kind of punishment on large-scale cooperation (Figure 7a,c), such evidence would have high value.

4. Open problems

We know now a lot more about human altruism than we did one decade ago. There is experimental evidence indicating that repeated interactions, reputation-formation, and strong reciprocity are powerful determinants of human behaviour. There are formal models which capture the subtleties of interactions between selfish and strongly reciprocal individuals and there is a much better understanding about the nature of the evolutionary forces that probably shaped human altruism. However, there are still a considerable number of open problems. In view of the relevance of cultural evolution, it is necessary to study the relationship between cultural and economic institutions and the prevailing patterns of human altruism. Although recent evidence (Henrich et al., 2001) suggests that market integration and the potential gains from cooperation are important factors, our knowledge is still extremely limited. This limitation is partly due to the fact that far too many experiments use students from advanced countries as participants. Instead, we need experiments with participants that are representative of whole countries or cultures and we need to combine behavioural measures of altruism with individual-level demographic data and group-level data about cultural and economic institutions. In view of the theoretical importance of group conflicts

and group reputation, much more evidence on how these affect altruistic rewarding and punishment is necessary. We also need more empirical knowledge about the characteristics of the individual reputation acquired by people and how others respond to this reputation. At the ultimate level, the evolution and role of altruistic rewarding for cooperation in larger groups remain in the dark. Likewise, the empirical study of altruistic rewarding has been largely limited to dyadic interactions and little is known about how cooperation in n-person public good situations is affected if subjects, after they observed each others' actions in the PD, have the opportunity to altruistically reward others. Evolutionary explanations of altruistic rewarding are likely to be much more difficult than explanations of altruistic punishment because, when cooperation is frequent, rewarding causes high costs for the altruists whereas a credible punishment threat renders actual punishment unnecessary. At the level of proximate theories of human motivation, we still lack parsimonious and tractable formal models of reciprocal fairness, which make precise, testable, predictions. Finally, to enhance the study of the evolution of altruism, there is a great need for sharp, empirically testable predictions that are rigorously derived from the evolutionary models.

Bibliography

Alexander, R. D. (1987), The Biology of Moral Systems. Aldine De Gruyter, New York

Andreoni, J./J. Miller (1993), Rational cooperation in the finitely repeated prisoner's dilemma: experimental evidence. Economic Journal, 103(418), 570-585

Andreoni, J. (1988), Why Free Ride? - Strategies and Learning in Public Goods Experiments. Journal of Public Economics, 37(3), 291-304

Andreoni, J. (1990), Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving. Economic Journal, 100(401), 464-477

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

Aoki, M. (1982), A condition for group selection to prevail over counteracting individual selection. Evolution, 36, 832-842

Axelrod, R./W. D. Hamilton (1981), The evolution of cooperation. Science, 211(4489), 1390-1396

Bellemare, C./S. Kröger (2003), On representative trust. Working Paper Tilburg University

Bendor, J./P. Swistak (2001), The evolution of norms. American Journal of Sociology, 106, 1493-1545

Berg, J./J. Dickhaut/K. McCabe (1995), Trust, Reciprocity and Social History. Games and Economic Behavior, 10(1), 122-142

Binmore, K. (Ed.) (1998), Just Playing. Game Theory and the Social Contract, 2. The MIT Press, Cambridge

Blount, S. (1995), When social outcomes aren't fair - the effect of causal attributions on preferences. Organizational Behavior and Human Decision Processes, 63(2), 131-144

Bolton, G. E./E. Katok/R. Zwick (1998), Dictator Game Giving: Rules of Fairness versus Acts of Kindness. International Journal of Game Theory, 27(2), 269-299

Bolton, G. E./A. Ockenfels (2000), ERC: A Theory of Equity, Reciprocity, and Com-

petition. American Economic Review, 90(1), 166-193
Bolton, G./R. Zwick (1995), Anonymity versus Punishment in Ultimatum Bargaining. Games and Economic Behavior, 10(1), 95-121
Bornstein, G./M. Ben-Yossef (1994), Cooperation in Intergroup and Single-Group Social Dilemmas. Journal of Experimental Social Psychology, 30(52-67)
Bornstein, G./I. Erev/H. Goren (1994), The effect of repeated play in the IPG and IPD team games. Journal of Conflict Resolution, 38(4), 690-707
Bowles, S./J.-K. Choi/A. Hopfensitz (2003), The Co-Evolution of Individual Behaviours and Social Institutions. Journal of Theoretical Biology, (in press)
Boyd, R./P. Richerson (2005), Solving the puzzle of human cooperation. In: S. Levinson and P. Jaisson (Eds.), Evolution and Culture. MIT Press, Cambridge MA
Boyd, R./H. Gintis/S. Bowles/P. J. Richerson (2003), The evolution of altruistic punishment. Proceedings of the National Academy of Sciences of the United States of America, 100(6), 3531-3535
Boyd, R./P. J. Richerson (1988), The evolution of reciprocity in sizable groups. Journal of Theoretical Biology, 132(3), 337-356
Boyd, R./P. J. Richerson (1992), Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethology and Sociobiology, 13(3), 171-195
Brandts, J./C. Sola (2001), Reference points and negative reciprocity in simple sequential games. Games and Economic Behavior, 36, 138-157
Buchan, N. R./R. T. A. Croson/R. M. Dawes (2002), Swift Neighbors and Persistent Strangers: A Cross-Cultural Investigation of Trust and Reciprocity in Social Exchange. American Journal of Sociology, 108(1), 168-206
Cameron, L. A. (1999), Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia. Economic Inquiry, 37(1), 47-59
Chaudhuri, A./S. Graziano (2003), Evolution of Conventions in an Experimantal Public Goods Game with Private and Public Knowledge of Advice. Working Paper
Cosmides, L./J. Tooby (1992), Cognitive Adaptations for Social Exchange. In: J. Barkow, L. Cosmides and J. Tooby (Eds.), The Adapted Mind. Oxford University Press, New York
DalBo, P. (2003), Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. Working Paper, Dept. of Econ., Brown University
Daly, M./M. Wilson (1988), Evolutionary Social-Psychology and Family Homicide. Science, 242(4878), 519-524
Dawes, R. M. (1980), Social Dilemmas. Annual Review of Psychology, 31, 169-193
de Quervain, D. J. F./U. Fischbacher/V. Treyer/M. Schelthammer/U. Schnyder et al. (2004), The neural basis of altruistic punishment. Science, 305(5688), 1254-1258
Dufwenberg, M./G. Kirchsteiger (2004), A theory of sequential reciprocity. Games and Economic Behavior, 47, 268-298
Ekman, P./M. O'Sullivan (1991), Who Can Catch a Liar? American Psychologist, 49(9), 913-920
Engelmann, D./U. Fischbacher (2002), Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game. Working Paper No. 132, Institute for Empirical Research in Economics, University of Zurich
Falk, A./U. Fischbacher (in press), A Theory of Reciprocity. Games and Economic Behavior
Falk, A./E. Fehr/U. Fischbacher (2003), On the nature of fair behavior. Economic Inquiry, 41(1), 20-26
Falk, A./E. Fehr/U. Fischbacher (in press), Driving forces of informal sanctions. Econometrica

Fehr, E./J. Henrich (2003), Is Strong Reciprocity a Maladaptation: On the Evolutionary Foundations of Human Altruism. In: P. Hammerstein (Ed.), Genetic and Cultural Evolution of Cooperation. Dahlem Workshop Report 90. The MIT Press, Cambridge

Fehr, E./E. Tougareva/U. Fischbacher (2002a), Do high stakes and competition undermine fairness? Working Paper No. 125, Institute for Empirical Research in Economics, University of Zurich

Fehr, E./U. Fischbacher (2003), The nature of human altruism. Nature, 425(6960), 785-791

Fehr, E./U. Fischbacher (2004), Third-party punishment and social norms. Evolution and Human Behavior, 25(2), 63-87

Fehr, E./U. Fischbacher/S. Gächter (2002b), Strong reciprocity, human cooperation, and the enforcement of social norms. Human Nature, 13(1), 1-25

Fehr, E./S. Gächter (2002), Altruistic punishment in humans. Nature, 415, 137-140

Fehr, E./S. Gächter/G. Kirchsteiger (1997), Reciprocity as a Contract Enforcement Device - Experimental Evidence. Econometrica, 65(4), 833-860

Fehr, E./G. Kirchsteiger/A. Riedl (1993), Does Fairness prevent Market Clearing? An Experimental Investigation. Quarterly Journal of Economics, 108(2), 437-459

Fehr, E./K. M. Schmidt (1999), A theory of fairness, competition, and cooperation. Quarterly Journal of Economics, 114(3), 817-868

Fershtman, C./U. Gneezy (2001), Discrimination in a Segmented Society: An Experimental Approach. Quarterly Journal of Economics, 115, 351-377

Festinger, L. (1954), A Theory of Social Comparison Processes. Human Relations, 7, 117-140

Fischbacher, U. (1998), Z-tree. Zurich toolbox for readymade economic experiments. Working Paper No. 21, Institute for Empirical Research in Economics, University of Zurich

Fischbacher, U./S. Gächter/E. Fehr (2001), Are people conditionally cooperative? Evidence from a public goods experiment. Economic Letters, 71, 197-404

Fischbacher, U./C. Fong/E. Fehr (2002), Fairness and the Power of Competition. Working Paper No. 133, Institute for Empirical Research in Economics, University of Zurich

Forsythe, R./J. L. Horowitz/N. E. Savin/M. Sefton (1994), Fairness in Simple Bargaining Experiments. Games and Economic Behavior, 6(3), 347-369

Frank, M. G./P. Ekman (1997), The Ability to Detect Deceit Generalizes Across Different Types of High-Stake Lies. Journal of Personality and Social Psychology, 72(6), 1429-1439

Frank, R. (1988), Passions within Reason. The Strategic Role of the Emotions. W.W. Norton & Company, New York

Frank, R. H./T. Gilovich/D. T. Regan (1993), The Evolution of One-Shot Cooperation: An Experiment. Ethology and Sociobiology, 14, 247-256

Frohlich, N./J. Oppenheimer (1984), Beyond Economic Man - Altruism Egalitarianism, and Difference Maximizing. Journal of Conflict Resolution, 28(1), 3-24

Gachter, S./A. Falk (2002), Reputation and reciprocity: consequences for the labour relation. Scandinavian Journal of Economics, 104(1), 1-26

Gintis, H. (2000), Strong reciprocity and human sociality. Journal of Theoretical Biology, 206, 169-179

Gintis, H./E. A. Smith/S. Bowles (2001), Costly signaling and cooperation. Journal of Theoretical Biology, 213(1), 103-119

Gintis, H. (2003), The Hitchkiker's Guide to Altruism: Gene-Culture Co-Evolution

and the Internalization of Norms. Journal of Theoretical Biology, 220, 407-418

Güth, W./R. Schmittberger/B. Schwarze (1982), An Experimental Analysis of Ultimatum Bargaining. Journal of Economic Behavior & Organization, 3, 367-388

Güth, W./R. Schmittberger/R. Tietz (1990), Ultimatum Bargaining Bahavior - A Survey and Comparison of Experimental Results. Journal of Economic Psychology, 11, 417-449

Hammerstein, P. (2003), Why Is Reciprocity So Rare in Social Animals. In: P. Hammerstein (Ed.), Genetic and Cultural Evolution of Cooperation. Dahlem Workshop Report 90. The MIT Press, Cambridge

Harbaugh, W. T./K. Krause/S. Liday (2000), Children's Bargaining Behavior: Differences by Age, Gender, and Height. Working Paper

Hayashi, N./E. Ostrom/J. Walker/T. Yamagishi (1999), Reciprocity, trust, and the sense of control - A cross-societal study. Rationality and Society, 11(1), 27-46

Henrich, J./R. Boyd (2001), Why people punish defectors—weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. Journal of Theoretical Biology, 208(1), 79-89

Henrich, J./R. Boyd/S. Bowles/C. Camerer/E. Fehr et al. (2001), In search of homo economicus: behavioral experiments in 15 small-scale societies. American Economic Review, 91(2), 73-78

Hill, K. (2002), Altruistic Cooperation During Foraging by the Ache, and the evolved Human Predisposition to Cooperate. Human Nature, 13(1), 105-128

Hoffman, E./K. McCabe/K. Shachat/V. Smith (1994), Preferences, Property Rights and Anonymity in Bargaining Games. Games and Economic Behavior, 7(3), 346-380

Hoffman, E./K. McCabe/V. Smith (1996), On Expectations and the Monetary Stakes in Ultimatum Games. International Journal of Game Theory, 25(3), 289-301

Isaac, M. R./K. McCue/C. R. Plott (1985), Public Goods Provision in an Experimental Environment. Journal of Public Economics, 26, 51-74

Isaac, M. R./J. M. Walker (1988), Group Size Effects in Public Goods Provision: The Voluntary Contribution Mechanism. Quarterly Journal of Economics, 103(1), 179-199

Isaac, R. M./J. Walker/S. Thomas (1984), Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations. Public Choice, 43(2), 113-149

Johnson, D. P./P. Stopka/S. Knights (2003), The puzzle of human cooperation. Nature, 421, 911-912

Jorgensen, J. G. (1980), Western Indians: Comparative Environments, Languages, and Cultures of 172 Western American Indian Tribes. W. H. Freeman, San Francisco

Kaplan, H./J. Hill/J. Lancaster/A. M. Hurtado (2000), A theory of human life history evolution: Diet, intelligence, and longevity. Evolutionary Antropology, 9(4), 156-185

Kelly, R. C. (1985), The Nuer Conquest: The structure and Development of an Expansionist System. University of Michigan Press, Ann Arbor

Ledyard, J. (1995), Public Goods: A Survey of Experimental Research. In: John Kagel and Alvin Roth (Eds.), Handbook of Experimental Economics. Princeton University Press, 111-194

Leimar, O./P. Hammerstein (2001), Evolution of cooperation through indirect reciprocity. Proceedings of the Royal Society of London Series B-Biological Sciences, 268(1468), 745-753

Levine, D. K. (1998), Modeling altruism and spitefulness in experiments. Review of Economic Dynamics, 1(3), 593-622

Loewenstein, G. F./L. Thompson/M. H. Bazerman (1989), Social Utility and Decison Making in Interpersonal Contexts. Journal of Personality and Social Psychology, 57(3), 426-441

Long, J. C. (1986), The allelic Correlation Structure of Gainj and Kalam speaking Peoples and Interpretation of Wright's F-Statistics. Genetics, 112, 629-647

MacCrimmon, K. R./D. M. Messick (1976), A Framework for Social Motives. Behavioral Science, 21, 86-100

Margolis, H. (1982), Selfishnes, Altruism & Rationality: A Theory of Social Choice. The University of Chicago Press, Chicago

Messick, D./M. Brewer (1983), Solving Social Dilemmas: A Review. In: L. Wheeler (Ed.), Review of Personality and Social Psychology. Sage Publications, Beverly Hills

Messick, D./K. Sentis (1985), Estimating social and nonsocial utility functions from ordinal data. European Journal of Social Psychology, 15, 389-399

Milinski, M./D. Semmann/T. C. M. Bakker/H. J. Krambeck (2001), Cooperation through indirect reciprocity: image scoring or standing strategy? Proceedings of the Royal Society of London Series B-Biological Sciences, 268(1484), 2495-2501

Milinski, M./D. Semmann/H. J. Krambeck (2002a), Donors to charity gain in both indirect reciprocity and political reputation. Proceedings of the Royal Society of London Series B-Biological Sciences, 269(1494), 881-883

Milinski, M./D. Semmann/H. J. Krambeck (2002b), Reputation helps solve the 'tragedy of the commons'. Nature, 415(6870), 424-426

Nowak, M. A./K. Sigmund (1998a), The dynamics of indirect reciprocity. Journal of Theoretical Biology, 194(4), 561-574

Nowak, M. A./K. Sigmund (1998b), Evolution of indirect reciprocity by image scoring. Nature, 393(6685), 573-577

Nowak, M./K. Sigmund (1993), A strategy of win stay, lose shift that outperforms tit-for-tat in the prisoners dilemma game. Nature, 364(6432), 56-58

Ostrom, E./J. Walker/R. Gardner (1992), Covenants With and Without a Sword: Self-Governance is Possible. American Political Science Review, 86, 404-417

Otterbein, K. F. (1985), The Evolution of War: A Cross-Cultural Study. Human Relations Area Files Press, New Haven

Panchanathan, K./R. Boyd (2004), Indirect reciprocity can stabilize cooperation without the second-order free rider problem. Nature

Rabbie, J. M./J. C. Schot/L. Visser (1989), Social Identity Theory: A Conceptual and Empirical Critique from the Perspective of a Behavioural Interaction Model. European Journal of Social Psychology, 19, 171-202

Rabin, M. (1993), Incorporating fairness into game theory and economics. American Economic Review, 83(5), 1281-1302

Rillling, J. K./D. A. Gutman/T. R. Zeh/G. Pagnoni/G. S. Berns et al. (2002), A neural basis for social cooperation. Neuron, 35, 395-405

Robson, A. (1990), Efficiency in Evolutionary Games: Darwin, Nash and Secret Handshake. Journal of Theoretical Biology, 144, 379-396

Rogers, A. R. (1990), Group selection by selective emigration: The effects of migration and kin structure. American Naturalist, 135, 398-413

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

Runciman, W. G. (1966), Relative Deprivation and Social Justice. Penguin, NY

Schotter, A. (2003), Decision Making with Naive Advice. American Economic Review, 93, 196-201

Seinen, I./A. Schram (in press), Social Status and Group Norms: Indirect Reciprocity in a Helping Experiment. European Economic Review

Silk, J. B. (1980), Adoption and kinship in oceania. American Anthropologist, 82, 799-820

Singer, T./S. J. Kiebel/J. S. Winston/H. Kaube/R. J. Dolan et al. (2004), Brain responses to the acquired moral status of faces. Neuron, 41(4), 653-662

Slonim, R./A. E. Roth (1998), Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic. Econometrica, 66(3), 569-596

Sober, E./D. S. Wilson (1998), Unto Others - The Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge, Massachusetts

Soltis, J./R. Boyd/P. J. Richerson (1995), Can Group-Functional Behaviors Evolve by Cultural-Group Selection - an Empirical-Test. Current Anthropology, 36(3), 473-494

Tajfel, H. (1982), Social Psychology of Intergroup Relations. Annual Review of Psychology, 33, 1-30

Tajfel, H./M. Billig/R. Bundy/C. Flament (1971), Social Categorization in Intergroup Behaviour. European Journal of Social Psychology, 1, 149-178

Thibaut, J. W./H. H. Kelley (1959), The Social Psychology of Groups. Wiley, New York

Trivers, R. L. (1971), Evolution of Reciprocal Altruism. Quarterly Review of Biology, 46(1), 35-57

Wedekind, C./M. Milinski (2000), Cooperation through image scoring in humans. Science, 288(5467), 850-852

Williams, G. D. (1966), Adaption and Natural Selection: A Critique of Some Current Evolutionary Thought. Princeton University Press, Princeton

Yamagishi, T./N. Jin/T. Kiyonari (1999), Bounded Generalized Reciprocity: Ingroup Boasting and Ingroup Favoritism. In: Shane R. Thye, Edward J. Lawler, Michael W. Macy and Henry A. Walker (Eds.), Advances in Group Processes. JAI Press, Stamford, 161-197

Yamagishi, T. (1986), The provision of a sanctioning system as a public good. Journal of Personality and Social Psychology, 51(1), 110-116

Yamagishi, T./T. Kiyonari (2000), The Group as the Container of Generalized Reciprocity. Social Psychology Quarterly, 63(2), 116-132

Zahavi, A. (1995), Altruism as a Handicap - the Limitations of Kin Selection and Reciprocity. Journal of Avian Biology, 26(1), 1-3

Analyse & Kritik 27/2005 (© Lucius & Lucius, Stuttgart) p. 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 and Tversky' 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 con-

tent 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 contractual 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 wages, job ladders, and the use of promotion and dismissal as motivating devices)

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

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