Cooperation in Primates and Humans

Mechanisms and Evolution

Edited by

Peter M. Kappeler & Carel P. van Schaik

Springer
Heidelberg
2005

Contents

Chapter 15

Human cooperation from an economic perspective

Simon Gächter, Benedikt Herrmann

15.1 Introduction

Many important economic and social situations are characterized by a conflict of interest between individual and group benefits. The 'tragedy of the commons' (Hardin 1968) is probably one of the best known examples of this problem. Each individual farmer has an incentive to put as many cattle on the common meadow as possible. The tragic consequence may be overgrazing from which all farmers suffer. Collectively, all farmers would be better off if they were able to constrain the number of cattle that simultaneously graze on the commons. Yet, each individual farmer is better off by letting his cattle graze. A similar tension between individual and collective rationality is typical in such diverse areas like warfare, cooperative hunting and foraging, environmental protection, tax compliance, voting, the participation in collective actions like demonstrations and strikes, the voluntary provision of public goods, donations to charities, teamwork, collusion between firms, embargos and consumer boycotts, and so on.

While the logic of self-interest is straightforward, the facts seem to be at odds with theoretical predictions derived under the joint assumptions of rationality and selfishness. At the societal level, our societies have achieved a degree of cooperation and division of labor among genetically unrelated individuals that is unprecedented in the animal kingdom (see Seabright 2004 for a recent account). At a lower level, the fact that people even in anonymous situations vote, take part in collective actions, often manage not to overuse common resources, care for the environment, mostly do not evade taxes on a large scale, donate to public radio, as well as to charities, etc. suggests that the strict self-interest hypothesis is inconsistent with the degree of cooperation that we observe around us.

How can we explain this? This paper presents evidence from systematic experimental investigations on how people solve cooperation problems. Laboratory experiments are probably the best tool for studying cooperation. The reason is that in the field many factors are operative at the same time. The laboratory allows for a degree of control that is not feasible in the field. In all the laboratory experiments that we will discuss below participants, depending on their decisions, earned considerable amounts of money. Thus, the laboratory allows observing real economic behavior under controlled circumstances (see Friedman & Sunder 1994 for an introduction to methods in experimental economics and Kagel & Roth 1995 for an overview of important results).

In the next section, we will introduce two prototypical cooperation games that have been extensively investigated in experiments: (i) the 'Prisoner's Dilemma' (PD) and (ii) the 'public goods experiment'. These games are simple and contain the essence of the cooperation problems introduced above. Many of them are structured such that purely selfish individuals would not cooperate in these games. Yet, we will show that there is substantial cooperation even in completely anonymous one-shot situations. This finding has been termed 'altruistic cooperation' or 'altruistic rewarding' (e.g. Fehr & Fischbacher 2003) because apparently some people are prepared to benefit others by cooperating. Yet, most of this 'altruistic cooperation' takes the form of 'conditional cooperation'; people cooperate if others cooperate as well. 'Altruistic rewarding' has also been observed in other contexts (for surveys, see Fehr & Gächter 2000a and Camerer 2003, chapter 2).

One of the most important insights from the laboratory experiments is that in the absence of extrinsic incentives like reputation, social (dis-)approval and punishment, cooperation is fragile. Cooperation almost inevitably breaks down in repeated interactions. The reason is that conditional cooperators can only avoid being exploited by the free riders if they stop cooperating themselves. The lack of targeted punishment leaves the cooperators with the only option they have, stopping cooperation.

In Section 15.3, we will look at reputation, communication and social approval. These important mechanisms are frequently available in reality and may help to sustain cooperation. Reputation mechanisms have recently gained a lot of attention. It turns out that reputation can have a strong cooperation-enhancing effect. The same holds for communication. Similarly, there is also experimental evidence that social approval can lead to a substantial increase in cooperation.

Section 15.4 presents evidence that shows that many people are prepared to engage in altruistic cooperation but also in 'altruistic punishment'. They do this even in anonymous one-shot situations in which future benefits from reciprocal altruism (Trivers 1971), indirect reciprocity and reputation (Alexander 1987, Nowak & Sigmund 1998), signaling (Zahavi & Zahavi 1997, Gintis et al. 2001) and kinship (Hamilton 1964) are excluded by the experimental design. This punishment is altruistic, because it is costly to the individual and beneficial for someone else who interacts with the punished (and now well-behaved) individual in the future.

Section 15.5 discusses the role of emotions as a proximate mechanism that can explain altruistic punishment. Section 15.6 looks at evolutionary explanations for the observed behavior. Section 15.7 presents a summary and some concluding remarks.

15.2 Some stylized facts on cooperation

We start our discussion with a brief presentation about what is known about factors influencing cooperation and free riding. The most important vehicles for

Table 15.1. The Prisoner's Dilemma. The amounts in each cell refer to the players' payoff. In each cell, the left payoff refers to player 1's payoff, and the right payoff to player 2's payoff.

		Player 2	
		Cooperate	**Defect**
Player 1	Cooperate	€80, €80	€0, €100
	Defect	€100, €0	€35, €35

studying cooperation problems in controlled laboratory experiments are the PD and the 'public goods experiment'.

Table 15.1 illustrates the prototype cooperation game, the famous PD (see Poundstone 1992 for an illuminating discussion of this game). In the game of Table 15.1, two players are told that they can choose simultaneously between 'Cooperate' and 'Defect'. If both choose 'Cooperate', both earn 80 Euros each. If player 1, for instance, chooses 'Defect', while player 2 chooses 'Cooperate', player 1 earns 100 Euros, while player 2 gets nothing. If player 1 cooperates and player 2 defects, player 1 will earn nothing and player 2 will earn 100 Euros. If both players defect, they earn 35 Euros each.

If this game is played only once, selfishness predicts no cooperation. Yet, if 'the shadow of the future' is important, i.e, if players interact for an unknown length of time, and if people are not too impatient and therefore care for the future, then strategic cooperation becomes possible, because defection can be punished by withholding future cooperation and even more complicated punishment strategies (e.g. Fudenberg & Maskin 1986). The most famous idea is probably reciprocal altruism (Trivers 1971) and the related strategy of 'tit-for-tat', which turned out to be a very successful strategy in an 'evolutionary contest' where strategies played against each other in a computer simulation (Axelrod & Hamilton 1981). Its essence is the idea that favors are reciprocated ("I'll scratch your back if you'll scratch mine") and that unhelpful behavior is reciprocated by withholding future help.

Yet, the assumption that players are forced together for an unknown number of interactions may not hold (see Hammerstein 2003b for an extensive critique). In reality, people stop interacting with disliked partners and change social groups. Moreover, throughout (evolutionary) history, social groups were frequently disbanded by warfare, famine and other catastrophes (e.g. Knauft 1991, Gintis 2000, Fehr & Henrich 2003). These arguments suggest that studying short-term cooperation games is worthwhile. Moreover, though highly insightful, the studies by Axelrod & Hamilton (1981) are not about real behavior but are computer simulations. Therefore, we will turn next to some selected behavioral evidence on cooperation in finite PD games.

The PD game is probably one of the most extensively investigated games (see Rapoport & Chammah 1965, Colman 1999 and Ledyard 1995 for overviews on the

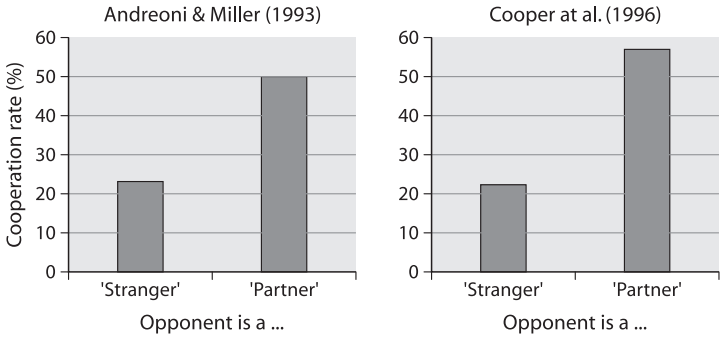

Fig. 15.1. Cooperation rates in the Prisoner's Dilemma. The figure shows the average cooperation rates from two studies, by Andreoni & Miller (1993) and Cooper et al. (1996), where players interacted for 10 periods, either with the same opponent ('Partner') or a randomly-matched opponent ('Stranger'). The prediction in both set-ups is a zero cooperation rate. Yet, in both set-ups, people cooperate, but substantially more in the 'Partner' than in the 'Stranger' set-up.

experimental evidence). Fig. 15.1 illustrates the results of two studies (by Cooper et al. 1996 and Andreoni & Miller 1993) in each of which the subjects played the game 10 times under two different conditions. In one condition, called the 'Stranger' condition, each player was matched with a new player in each of the 10 periods. In the second condition, the 'Partner' condition, the opponent stayed the same throughout all repetitions of the game. The subjects were informed about this. Thus, under the assumption of selfishness and rationality, all players in both conditions are predicted to defect. In the 'Stranger' condition, this prediction holds because each play of the game is against a new opponent and hence 'one-shot'. In the 'Partner' condition, the prediction holds with backward induction; in the last period, both players (who are assumed to be rational and selfish) will defect. Therefore, in the penultimate period, there is no incentive to cooperate, since players will surely defect in the last period. Hence, there is also no incentive to cooperate in the period prior to the penultimate one. Continuing this logic further implies that rational and selfish players will defect throughout. By contrast, if people are not completely sure that everyone is selfish, then it might pay to build up a reputation by cooperating if others cooperate until the final rounds, where a selfish player should defect for sure (see Kreps et al. 1982 for a game-theoretical explanation and Selten & Stoecker 1986 for a bounded rationality approach).

In both studies, the results in the 'Stranger' condition are at odds with this prediction. People cooperate on average in slightly more than 20% of the cases. To have a common future, if only for 10 rounds increases cooperation substantially. In the 'Partner' condition, the average cooperation rate is at least 50%. Thus: (i) people are prepared to cooperate even in one-shot games and (ii) the possibility of behaving strategically strongly increases cooperation.

Clark & Sefton (2001) studied an interesting variation of the game of Table 15.1. Instead of playing the game simultaneously, their subjects played the game sequentially, i.e. player 1 first made his or her choice, which was then observed

by player 2 before deciding whether to cooperate or to defect. The subjects also played the game for 10 rounds in the 'Stranger' set-up. Clark & Sefton (2001) find that between 37% and 42% of the subjects cooperate conditionally on others' cooperation. Such conditional cooperation is also observed in two further treatments, 'double temptation', where the defection payoff was doubled, and 'double stakes', in which all payoffs were doubled. A statistical analysis shows that under 'double temptation', the fraction of conditional cooperation is reduced relative to the baseline, whereas 'double stakes' did not significantly affect the extent of conditional cooperation. Experiments on the sequential PD where the two players could also choose intermediate cooperation levels confirm the importance of conditional cooperation (e.g. Fehr et al. 1993, Fehr et al. 1997, Falk et al. 1999, Gächter & Falk 2002; see Fehr & Gächter 2000a for an overview).

These results are interesting, because the PD is such a simple and generic cooperation game. The fact that people cooperate (conditionally) even in one-shot games casts doubt on the selfishness assumption. The observation that there are strong effects of repeated interaction suggests that straightforward economic incentives are very helpful for successful cooperation. There can thus be no doubt that reciprocal altruism and the strategic gains from cooperation that come from repeated interactions are a powerful force in explaining real-world cooperation in small and stable groups. Yet, the success of reciprocal altruism in sustaining cooperation may be limited if groups become bigger. As has been shown theoretically (see Boyd & Richerson 1988), cooperation in the PD can only be sustained in groups larger than $n > 2$ if all other group members cooperated in the previous period. Thus, the basin of attraction for cooperation is very small because a few free riders can undermine cooperation. For this theoretical reason, it is worthwhile to move beyond dyadic relationships.

The most commonly used game for studying n-person cooperation problems is the public goods game. In contrast to a private good, a public good is a good which can be consumed even if one has not paid for it, or not contributed to its provision. Clean air, environmental quality and national security, but also collective reputations or team output are common examples of public goods.

An economic model of public goods provision is the public goods game. This game underlies many experiments that study cooperation for the provision of public goods. In a typical public goods experiment, four people form a group. All group members are endowed with 20 tokens. Each subject *i* has to decide independently how many tokens (between zero and 20) to contribute to a common project (the public good). The contributions of the whole group are summed up. The experimenter then multiplies the sum of contributions by 1.6 and distributes the resulting amount equally among the four group members. Thus each subject *i*'s payoff is

$$\pi_i = 20 - g_i + \frac{1.6}{4} \sum_{j=1}^{4} g_j$$

The first term $(20 - g_i)$ indicates the payoff from the tokens not contributed to the public good (the 'private payoff'). The second term is the payoff from the public good. Each token contributed to the public good becomes worth 1.6 to-

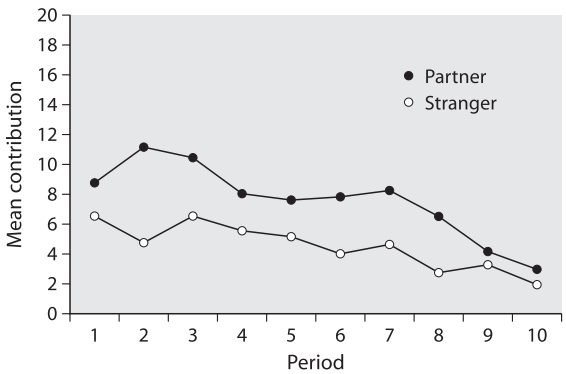

Fig. 15.2. Contributions to a public good in constant ('Partner') and randomly-changing groups ('Strangers') over 10 repetitions. Cooperation gains are maximized with full contributions (20 tokens). Selfishness predicts zero contributions. The figure shows that 'Partners' contribute more than 'Strangers' and that cooperation collapses in both treatments. From Fehr & Gächter (2000b).

kens. The resulting amount is distributed equally among the four group members, irrespective how much an individual has contributed. Thus, an individual benefits from the contributions of other group members, even if he or she has contributed nothing to the public good. Therefore, a rational and selfish individual has an incentive to keep all tokens for him- or herself, since the 'return' per token from the public good for him- or herself is only 0.4 (1.6/4), whereas it is one if he or she keeps the token. By contrast, the group as a whole is best off if everybody contributes all 20 tokens.

Since the public goods game is an n-person cooperation problem that is easy to implement and since it also reflects the tension between individual incentives and collective benefits, it has been frequently used in experimental studies (see Ledyard 1995 for an overview). Fig. 15.2 depicts a typical finding of a public goods experiment, where the exact same game is repeated 10 times and subjects know this. In each period, subjects receive 20 tokens and decide how many of them to keep or contribute to the public good. After each round, subjects are informed about what the other group members have contributed. Fig. 15.2 shows the resulting cooperation patterns in a 'Stranger' condition, where group members change randomly from round to round, and a 'Partner' condition, in which groups stay constant for all rounds.

Look at the 'Strangers' data first. Mean contributions start at about 6.5 tokens and decline to about two tokens in the 10 iterations of the public good game. In other words, by the end of the experiment, cooperation has almost entirely collapsed. As in the repeated PD, we find that cooperation in the 'Partner' condition is higher from the very beginning of the experiment. Yet, by the tenth round, cooperation has collapsed as well.

Fig. 15.2 illustrates two stylized facts from dozens of public goods experiments. First, as in the PD experiments reported above, 'Partners' contribute more than 'Strangers' (see Keser & van Winden 2000, and Andreoni & Croson 1998 for an overview). This result has also been found in other cooperation games (e.g.

Falk et al. 1999, Gächter & Falk 2002, Fehr & Fischbacher 2003). The significance of this and related findings is that people are immediately able to distinguish whether they are in a situation that requires strategic cooperation (the 'Partner' condition) or not (the 'Stranger' condition) and to adopt their behavior accordingly.

A second stylized fact is that cooperation is very fragile and tends to collapse with repeated interactions. Why is this so? One explanation is that people have to learn how to play this game. Since errors can only go in one direction, any erroneous decision looks like a contribution. Over time, people learn and commit fewer errors, which is why contributions decline (Palfrey & Prisbrey 1997). The problem with this explanation is that it is inconsistent with the fact that after a so-called 'restart' (after the tenth round, participants are told that they will play another 10 rounds), cooperation jumps up again and basically starts at the same level as in the first period. If learning would explain the decay in cooperation, then, after the restart, cooperation should have continued at the level at which cooperation was in the tenth round (see Andreoni 1988). A second explanation is that people are heterogeneous with respect to their cooperative inclinations. Some people are free riders who try to maximize their monetary income, irrespective of other group members' contribution. Other people are 'conditional cooperators', who cooperate if others cooperate.

To test this idea, Fischbacher et al. (2001) invented a design that allows measuring the 'type' of a player by observing each participant's contribution to the public good as a function of other group members' contributions. Specifically, subjects were asked to indicate for each possible average contribution of the other group members how much they would like to contribute to the public good. The payoff function is the same as in the other public goods experiments; i.e., incentives are such that, given others' average contribution, the monetary income is always highest if one contributes nothing. Thus, a free rider type will always contribute zero to the public good. A conditional cooperator type will increase his or her contribution in the average contribution of others.

Fig. 15.3 shows the results of experiments that applied the Fischbacher et al. (2001) design. Fig. 15.3 contains the pooled results of the experiments by Fischbacher et al. (2001) and Fischbacher & Gächter (2004) who conducted their experiments in Switzerland with n = 44 and n = 140 subjects, respectively. The authors also ran experiments with n = 148 subjects in various cities in Russia.

24% percent of the Swiss subjects turned out to be free riders who contribute nothing for all contributions of the other group members. In our Russian subject pools, this frequency is markedly lower. Only 7% turned out to be free riders. By contrast, the fraction of conditional cooperators who cooperate if others cooperate is strikingly similar in Russia and Switzerland. In Switzerland, 54% of the subjects show contributions that increase in others' contribution, whereas in Russia this is true for 57%. Fig. 15.3 shows the average contribution of all conditional cooperators. We find that not only the fraction of conditional cooperators, but also the average contribution schedules are very similar. The only difference is that our Russian subjects are prepared to contribute slightly more for a given contribution of the other group members than the Swiss subjects. A further remarkable result is that the average contribution schedule of conditional

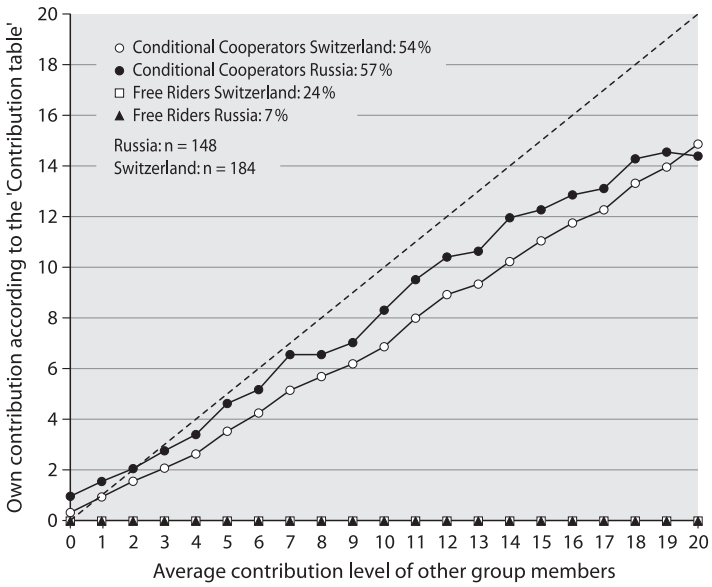

Fig. 15.3. The figure shows the mean contributions of different types of players to the public good as a function of other group members' average contributions. Free riders contribute nothing to the public good, irrespective of how much other group members contribute. In Switzerland [Russia], 24% [7%] of the subjects were free riders. Conditional cooperators increase their contributions the more others contribute. The graph 'Conditional Cooperators' is the average contribution of all subjects who report a contribution pattern that is increasing in other group members' contribution. In Switzerland [Russia], 54% [57%] of the subjects were conditional cooperators. From Fischbacher et al. (2001), Fischbacher & Gächter (2004) and new data from various places in Russia.

cooperators is 'self-servingly biased' because it is below the diagonal. Although conditional cooperators increase their contribution in the average contribution of the other group members, they do not fully match others' contribution.

How can the heterogeneity of types explain the fragility of cooperation that is so typical of repeatedly-played cooperation experiments (see Fig. 15.2)? The idea is simple. Conditional cooperators are prepared to cooperate if others cooperate. If they realize that others are taking a free ride, they reduce their contribution because they do not want to be 'suckered'. Moreover, even conditional cooperators have a 'self-serving bias'. Therefore, cooperation is bound to be fragile, even if most people are conditional cooperators (see Fischbacher & Gächter 2004 for a rigorous analysis).

Cooperation is even fragile if there is a leader who first decides on the contribution to the public good (e.g. Moxnes & van der Heijden 2003, Gächter & Renner 2004, Güth et al. 2004). This is remarkable, since one would expect that a leader should be able to utilize conditional cooperation by setting a good example. Yet, although conditional cooperation exists, free riding is there as well. Thus, the followers' cooperation is insufficient for inducing leaders to keep up their good example. Leaders get frustrated and stop setting a good example.

If it is indeed the mixture of types in a randomly-composed group that makes cooperation a fragile business, then an implication is that groups, where players know that others are of their type, should behave differently than randomly-composed groups. Specifically, conditional cooperators, who know that others are conditional cooperators as well, should find it easy to cooperate. To test this idea, Gächter & Thöni (in prep.) first had subjects play a one-shot public goods game. Then new groups were formed on the basis of the contribution to the public good in the one-shot game. The top cooperators were put in one group, the second to top in the next group and so on. After people had been sorted into the new groups, they were informed about this mechanism. Then they played the public goods experiment as 'Partners' for 10 rounds. It turned out that cooperators who knew that they were among other 'like-minded' cooperators, were able to maintain almost full cooperation until the final rounds. Surprisingly, even groups composed of free riders contributed to the public good. Yet, in stark contrast to the cooperator groups, cooperation among free riders entirely collapsed in the final period. Thus, they cooperated for purely strategic reasons and stopped doing so, when there was no future gain from cooperation anymore. The significance of this result is that the type composition and the knowledge of it (i.e. knowing that one is among like-minded players) matters strongly for the fragility of cooperation.

The experiments discussed so far looked at the most basic cooperation problem that exists in the absence of any extrinsic incentives, like reputation, social (dis-)approval and punishment. Any achieved cooperation must come from people's intrinsic readiness to cooperate, be it for strategic reasons and/or cooperative preferences. The results show that strategic incentives in a repeated interaction clearly help, but that cooperation is nevertheless fragile, with the exception of cooperators who know that they are among other like-minded cooperators. In the following, we look at evidence of how extrinsic incentives other than punishment mitigate the cooperation problem.

15.3 Reputation, communication and social approval

Humans often help each other or cooperate even if this act of altruism is not likely to be reciprocated. An important mechanism that may explain this kind of behavior in reality is reputation. One's behavior is often observed by third parties who may then decide to cooperate or not. Richard Alexander (1987) has coined the term 'indirect reciprocity' for such behavior, to distinguish it from direct reciprocity that occurs between two people. The idea is that helping someone, or refusing to help, changes ones social status, called 'image score'. People with a high image score are more likely to receive help from others: "Give and you shall receive". Game-theoretic analyses show that indirect reciprocity can be an evolutionary stable strategy (Nowak & Sigmund 1998). Seinen & Schram (in prep.), Engelmann & Fischbacher (2002) and Milinski and colleagues (see chapter 14) confirmed this experimentally; players with high image scores received more help than those with low image scores. Thus, indirect reciprocity is a mechanism

that can sustain cooperation even in situations in which direct reciprocity is not feasible. Milinski et al. (2002b also showed that a good reputation helps a player also in other social activities that involve the same partners (see Milinski, chapter 14 and Panchanathan & Boyd 2004 for a theoretical model).

In addition to the straightforward economic incentives conferred by reputation, people in reality also sometimes have the chance to communicate about their cooperation problem. To the extent that such communication does not lead to binding agreements, but is merely 'cheap talk', it might not necessarily help to solve the cooperation problem. A free rider might well promise to cooperate but then go on and defect, if his or her cooperation cannot be enforced. Therefore, theoretically, it is not at all clear why communication should reduce free riding. However, casual evidence and intuition suggest that communication has an impact. Thus, there are competing hypotheses and the lab may be the judge. Dawes et al. (1977) and Isaac & Walker (1988) were among the first to study the role of communication in cooperation. In the public goods experiments of Isaac & Walker (1988), group members could talk between the 10 rounds of the game. In a control treatment, communication was not possible. In this latter treatment, again, cooperation collapsed during repeat play. When face-to-face communication was possible, cooperation was substantially higher relative to the control treatment. Almost full efficiency, even in the final rounds, was achieved. Bochet et al. (in prep.) found that even anonymous 'chat-room' communication can lead to very high cooperation rates. Thus, communication can be a very powerful device for sustaining cooperation (see also Brosig et al. 2003 and Sally 1995 for an overview).

Yet, it is not entirely clear why exactly communication works. If many people are conditional cooperators, communication may help coordinating on a certain cooperation level. However, communication, in particular if it is face to face, is also a highly loaded psychological process that creates social ties and disseminates social (dis-)approval. People might fear the disapproval of others or might want to win their approval.

Rege & Telle (2004) developed a very simple one-shot experiment to test for social approval effects. In the control experiment, subjects simply made an anonymous contribution decision. In the main treatment, a subject's decision was publicly but silently recorded on a blackboard. All other participants could see the decision. From a standard economic viewpoint, this treatment manipulation should be ineffective. However, contributions were substantially higher when they could be observed than when they were anonymous.

Gächter & Fehr (1999) also tested the influence of social approval on cooperation. In their experiments, groups of four played the game for 10 repetitions as 'partners'. There were four treatments. In the benchmark 'Anonymity treatment', contributions and group members were anonymous throughout. In the 'Social Exchange treatment', group members were informed that they would get to know each other at the end and that then they would also learn each other's individual contributions during the game. In the 'Group Identity treatment', group members were introduced to each other before they played the game. Thus, a group identity could be formed. At the end of the experiment, they left the building individually such that they could not meet each other. Subjects were aware of

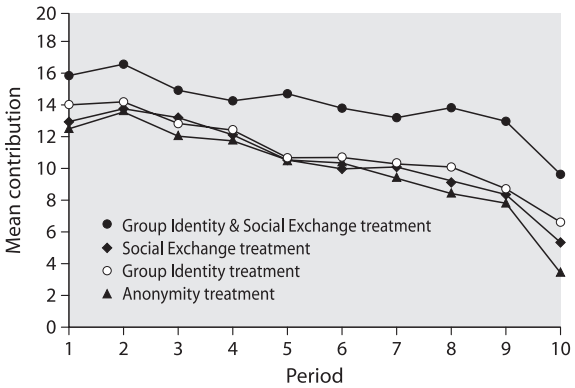

Fig. 15.4. The influence of group identity and social exchange on cooperation. The figure shows the mean contributions of 'Partners' to the public good. 'Group Identity' means that group members know each other's identity before they play; 'Social Exchange' means that subjects meet after the experiment to discuss what they did. In the 'Anonymity treatment', people neither meet before nor after the game. The figure shows that in this experiment only a combination of 'Group Identity' and 'Social Exchange' possibilities increases contributions. Cooperation is fragile in all treatments. From Gächter & Fehr (1999).

this. In the 'Group Identity & Social Exchange treatment', group members met before they played and where informed about the post-experimental meeting. In all four treatments, during the experiment subjects could not talk to each other. Between each round, they were only informed about the group average contribution. How can these treatments influence cooperation? Social exchange theory (e.g. Blau 1964) argues that people might exchange cooperation for social approval. Therefore, if people anticipate social approval effects, cooperation might be higher than under anonymity. Likewise, as suggested by psychological theory and previous evidence, group identity might increase cooperation (see for example Dawes et al. 1988). Fig. 15.4 shows the results.

The results show that, contrary to the hypotheses, neither group identity, nor social exchange alone, were able to increase cooperation. Only if both group identity and social exchange were possible did cooperation increase substantially relative to the anonymity benchmark. This result is consistent with the findings of Rege & Telle (2004). Yet, Rege & Telle (2004) only played their game once. The results from Fig. 15.4 show that social exchange, even if it increases cooperation, is not able to break the downward trend in cooperation. Cooperation is still very fragile, albeit at a higher level.

In summary, under appropriate circumstances, there is no doubt that reciprocal altruism, indirect reciprocity and reputation, and communication as well as social approval can enhance cooperation. A reason for observing higher cooperation when social approval is possible might be that the threat of disapproval of known group members induces higher cooperation rates. Thus, disapproval works like punishment. In fact, the group discussions at the end of the social exchange treatments often revealed quite some anger and frustration

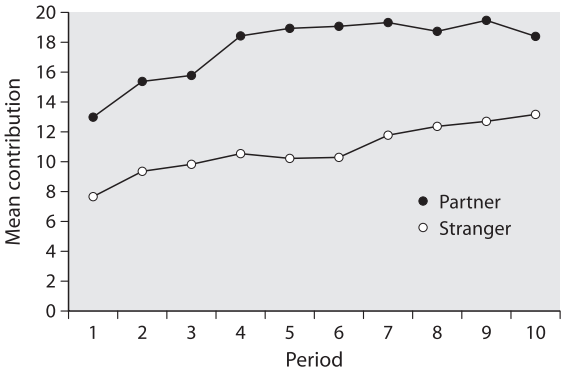

Fig. 15.5. Mean contributions to the public good in the presence of a punishment opportunity. The figure shows that contributions are substantially higher among 'Partners' than among 'Strangers'. A comparison with Fig. 15.2 shows that contributions to the public good are much higher and more stable when punishment is possible. From Fehr & Gächter (2000b).

towards the free riders. Since during the experiment 'social disapproval' could not be targeted at a free rider, it might not have been enough of a deterrent. In the next section, we therefore look at targeted punishment as a means to enhance cooperation.

15.4 Altruistic punishment and cooperation

Casual evidence as well as the observation reported above suggests that many people are in principle prepared to cooperate but want to avoid being the 'sucker' in social dilemma situations. Recall from Fig. 15.3 that roughly half of our subject pools are conditional cooperators who cooperate if others cooperate. If these people encounter a free rider in a typical anonymous standard public goods experiment, the only way to avoid being the 'sucker' is to withhold one's own cooperation. Since people typically strongly dislike being the 'sucker', they may be prepared to punish free riders if they could target them individually and even if it were costly for the punisher.

Yamagishi (1986) and Ostrom et al. (1992) were among the first to allow for punishment in interesting games. Yamagishi (1986) looked at people's willingness to provide a sanctioning system that itself is a public good. Ostrom et al. (1992) studied punishment in a common pool extraction system. Yet, these studies were not primarily interested in how people punish free riders. This was the focus of Fehr & Gächter (2000b) who developed an experimental design that allowed studying punishment in a public goods game. Specifically, after subjects had made their contributions to the public good, they entered a second stage, where they were informed about each individual group member's contribution. They could then assign up to 10 punishment points to each individual group

member. Punishment was costly for the punishing subject and each punishment point received reduced the punished subject's income from the first stage by 10%. Fehr & Gächter (2000b) played this experiment under two treatment conditions, a 'Partner'-treatment, where group members knew that they would play the game with the same four group members for 10 periods, and the 'Stranger'-treatment, where group composition was changed from period to period. Fehr & Gächter (2000b) also ran control experiments in which punishment was not possible (see Fig. 15.2). Fig. 15.5 shows the results in the treatments with punishment.

As the comparison with Fig. 15.2 shows, contributions to the public good are strongly increased in the presence of a punishment opportunity. This is true for both the 'Partner'- and the 'Stranger'-treatment. In the case of the 'Partner'-treatment, contributions approach almost 100% of the endowment; in the 'Stranger'-treatment contributions amount to 60% of the endowment. Thus, again we see that 'Partners' contribute more than 'Strangers'. From the very first period onward, contributions are significantly higher in the 'Partner'-treatment than in the 'Stranger'-treatment.

A theoretically very important question concerns the relevance of future interactions. In the 'Partner'-treatment, the likelihood of future interaction is one; in the 'Stranger'-treatment, where groups are randomly re-matched, it is much smaller (depending on the size of the pool from which groups are re-matched), but still positive. An interesting benchmark case is the situation where the likelihood of future interaction is zero, i.e. groups play a one-shot game. This situation is interesting, because neither reciprocal altruism and indirect reciprocity, nor any other form of reputation building is possible, since they require some future interactions. Therefore, Fehr & Gächter (2002) set up a so-called 'Perfect Stranger' design where in each of the six repetitions all groups were composed of completely new members, and participants knew this. Subjects played both games with no punishment and games with punishment. Half of the subjects started with the no-punishment condition and then were introduced to the punishment condition. For the other half, this order was reversed. Fig. 15.6 contains the results on the cooperation rates achieved.

The results are very clear-cut. When punishment is not available, cooperation collapses, as in all previous experiments. The picture changes dramatically, when punishment is possible. For instance, in the experiments that started with the punishment option (labeled '1. punishment, 2. no punishment'), contributions in the very first period were significantly higher than in the experiment that started with the no punishment option. In the experiments where punishment was introduced in the second sequence, cooperation jumped up immediately. This is remarkable, because in this sequence subjects experienced a strong decline in the games with no punishment. Still, after punishment had been introduced, cooperation jumped up to a level that even exceeded cooperation in the very first period. In both sequences, cooperation in the presence of a punishment opportunity strongly increased over time. Thus, contrary to theoretical predictions, in the presence of punishment, cooperation can flourish even in purely one-shot interactions.

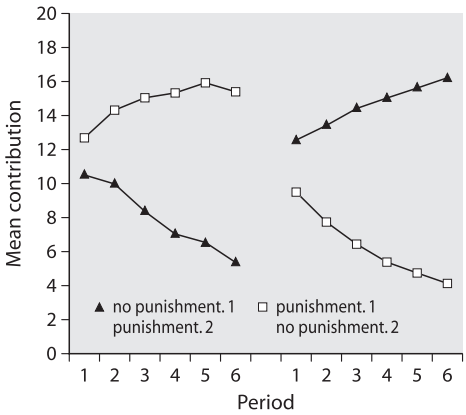

Fig. 15.6. Mean contributions to the public good among 'Perfect Strangers' in the absence and presence of a punishment option. In the sequence labeled "'1. no punishment, 2. punishment', subjects first played six rounds without the punishment option and were then introduced to an environment where they had a punishment option available in each of the following six rounds. In the sequence '1. punishment, 2. no punishment', subjects started in the game with punishment and were after the sixth round informed that there would be no punishment option in the next six rounds. The results show that contributions increase in the presence of punishment and decrease in its absence. From Fehr & Gächter (2002).

The reason why cooperation strongly increased in the presence of punishment is that cooperators were prepared to punish the free riders. Fig. 15.7 shows (separately for the 'Partner'-, the 'Stranger'- and the 'Perfect Stranger'-experiments in Zurich) the punishment expenditures for a given deviation from the other group members' average contributions. Fig. 15.7 also shows the punishment in a 'Partner'-experiment conducted in Samara (Russia). We will discuss this experiment below.

A couple of observations can be made from Fig. 15.7. First, the more a subject's contribution falls short of the average contribution of the other group members, the stronger is the punishment for the deviating group member. This is true in all treatments. Second, with the exception of very strong negative deviations (which comprise only a few cases, however) punishment is very similar between treatments. This is quite remarkable because cooperation levels differ strongly between the 'Partner', 'Stranger' and 'Perfect Stranger' treatments (compare Figs. 15.2, 15.5 and 15.6). In our view, this suggests that punishment is to a very large degree non-strategic. This view is also corroborated by the fact that the punishment pattern of Fig. 15.7 is temporally stable; i.e., some people are prepared to harm a free rider even in the final periods.

Why is punishment so successful in increasing cooperation? The most important reason is probably that it gives the selfish subjects, who care most about their individual payoff, a material incentive to cooperate. Since altruistic punishment is frequent, it apparently is a credible threat and induces selfish individuals to cooperate. It is exactly this feature that makes punishment altruistic;

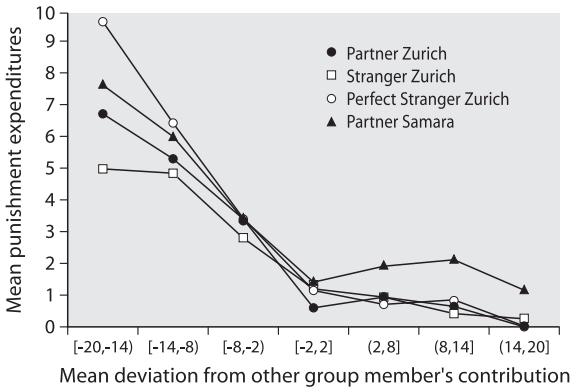

Fig. 15.7. Mean expenditures on punishment as a function of the deviation of the punished group member's cooperation from the average cooperation of the other members. The data are from experiments with 'Partners' and ('Perfect') 'Strangers' in Zurich and Samara. Each money unit spent on punishment reduced the income of the punished member by three money units. For example, group members spent 10 money units on punishing individuals whose contribution to the public good deviated between –20 and –14 units from the group average contribution. The data show that the more people free ride, the more altruistic punishment prevails. There is also some punishment of above-average contributors, in particular in the Samara subject pool. From Fehr & Gächter (2000b, 2002), Gächter et al. (2003).

a punished free rider might in his next encounter abstain from defecting, which benefits his or her future interaction partners.

By now, these results have been replicated by many researchers (see for example, Bowles et al. 2001, Sefton et al. 2002, Gächter et al. 2003, Masclet et al. 2003, Carpenter 2004, Carpenter et al. 2004, Falk et al. 2004, Gürerk et al. 2004, Anderson & Putterman, in prep., Bochet et al., in prep., Page et al. in prep., Carpenter in prep., Noussair & Tucker, in prep.). For lack of space, they cannot all be discussed here. We focus on three issues: (i) the perception of punishment, (ii) the demand for punishment, and (iii) cross-societal differences in norms of cooperation and punishment.

- **The perception of punishment.** A punishment may contain two messages. On the one hand, punishment directly inflicts a payoff reduction. On the other hand, punishment may also signal disapproval; i.e., it sends a message about socially inappropriate behavior. Both may be perceived as punishment and may therefore increase cooperation. Masclet et al. (2003) tested this intuition and studied 'formal and informal' sanctions. The structure of both formal and informal sanctions was the same as in Fehr & Gächter (2000b). Yet, while the formal sanctions were costly both for the punisher and the punished subjects, the informal sanctions were free; they neither caused costs for the punisher, nor the punished individual. Thus, they are tantamount to a symbolic disapproval. Consistent with the evidence on social approval effects reported above, it turned out that even informal sanctions were able to increase contributions. Yet, cooperation was more stable with formal than with informal punishment. In the experiments of Noussair & Tucker (in prep.), subjects

could use both formal and informal sanctions. It turned out that their combination led to higher contributions than either formal or informal sanctions alone.

- **The demand for punishment.** One of the most fundamental concepts in economics that underlies much of economic theory is the 'Law of Demand', according to which people will demand less of a certain commodity or activity the higher its price. Thus, from an economic viewpoint, an important question is whether this 'Law of Demand' also holds for punishment. Fig. 15.7 and all papers that have studied punishment in the context of a cooperation game confirm that many people do have a 'demand for punishment', in the sense that they are willing to pay a certain amount of money to inflict punishment on others (i.e. they 'buy' punishment). The more a subject free rides, the higher is the demand for punishment. Yet, studying the 'Law of Demand' requires a systematic variation of the cost of punishment. This is what Anderson & Putterman (in prep.) and Carpenter (2004) did. Their subjects played the cooperation and punishment game in the 'Stranger' set-up to minimize strategic effects. In each of the games, subjects faced different costs for inflicting a punishment unit on the punished subject. The results confirm that people demand less punishment, for a given amount of free riding, the higher the costs of punishing are. Thus, the 'Law of Demand' holds for punishment.
- **Cross-societal differences.** Cross-societal differences in norms of fair sharing have recently attracted a lot of attention (e.g. Henrich et al. 2001, Oosterbeek et al. 2004). It is therefore an interesting question to what extent there are differences in cooperation and punishment norms. To examine this question, Gächter et al. (2003) ran experiments in Russia, where they exactly replicated the Zurich 'Partner'-experiments. Fig. 15.7 also contains the punishment pattern for the Samara subjects. We find that the punishment of free riders is very similar to that in Zurich. Yet, above-average contributors in Samara experienced substantially more punishment than their counterparts in Zurich. Fig. 15.8 looks at the consequences of such punishment for cooperation behavior.

A comparison with the 'Partner'-experiments in Fig. 15.5 yields a striking difference, in particular when a punishment option is available. In the exact same experiment, the Zurich subjects were able to achieve almost full cooperation. By contrast, the presence of a punishment option is only able to prevent the collapse of cooperation. The average cooperation the Samara subjects achieve is only 68% of the level the Zurich subjects manage to maintain. Another stark difference is that in the Zurich experiments the presence of a punishment option strongly increased cooperation relative to cooperation in the absence of punishment (compare Figs. 15.2 and 15.5). This is not at all the case in Samara. Here, cooperation is not statistically significantly higher when subjects have a punishment option available. A potential explanation lies in the punishment behavior. As was shown in Fig. 15.7, the Samara subjects often substantially punished the above-average cooperators. This probably scared them off and thereby prevented the average cooperation level from increasing.

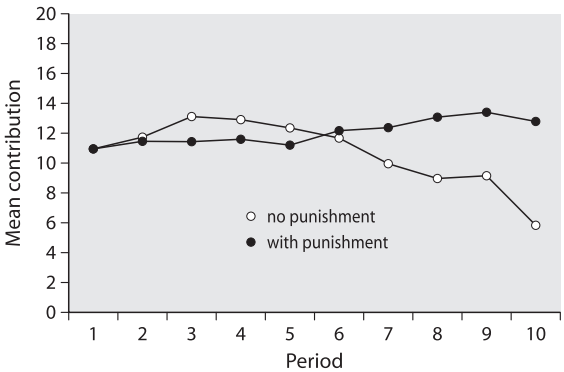

Fig. 15.8. The figure shows the mean contributions to the public goods in the absence and presence of punishment in a 10 times repeated 'Partner' experiment in Samara (Russia). In stark contrast to the results from Figs. 15.5 and 15.6, contributions are not significantly higher when punishment is possible. From Gächter et al. (2003).

In our view, the significance of this result is that different social groups may have widely differing social norms of cooperation, and in particular of punishment. This preliminary result suggests that it is worthwhile to understand logic and scope of cross-societal differences in norms of cooperation and punishment.

15.5 Emotions as a proximate mechanism

Given punishment, subjects' cooperation behavior looks quite rational. To avoid punishment, subjects cooperate in accordance with the group norm. Yet, why do people punish free riders in a one-shot context although this is costly? Emotions may play a decisive role here (Fessler & Haley 2003) and negative emotions, in particular, may provide a proximate explanation. Free riding may cause strong negative emotions among the cooperators and these emotions, in turn, may trigger the willingness to punish the free riders. If this conjecture is correct, we should observe particular emotional patterns in response to free riding. To elicit these patterns, the participants of the Fehr & Gächter (2002) experiments and the subjects in the Samara experiments were confronted with the following two hypothetical scenarios after the final period of the second treatment. The numbers in square brackets relate to the second scenario.

"You decide to invest 16 [5] francs to the project. The second group member invests 14 [3] and the third 18 [7] francs. Suppose the fourth member invests 2 [20] francs to the project. You now accidentally meet this member. Please indicate your feeling towards this person."

After they had read a scenario, subjects had to indicate the intensity of their anger and annoyance towards the fourth person (the free rider) on a seven-point

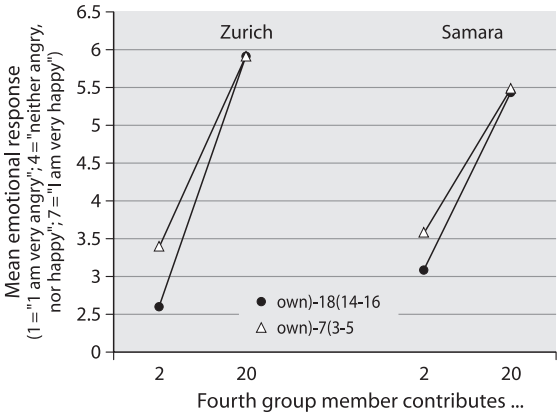

Fig. 15.9. Emotions as a proximate mechanism. The data are elicited in scenarios that describe own and others' contributions and then elicit one's own emotion toward a contribution of a 'fourth group member'. For instance, in Zurich, subjects who in the scenario contributed 16 tokens (whereas two others contributed 14 and 18 tokens, respectively) expressed an emotion score of 2.6 toward the fourth group member who only contributed 2 tokens. The emotion score is 5.9 if the fourth member contributes 20 tokens. The results show that people experience negative emotions toward a free rider more strongly the higher their own contribution level. The Samara subjects expressed less intensive emotions both toward the free rider and the high contributor. From Fehr & Gächter (2002; n = 240) and new results (n = 220).

scale (1 = 'very angry', 4 = 'neither angry nor happy', 7 = 'very happy'). The difference between Scenario 1 and 2 is that the other three persons in the group contribute relatively much in Scenario 1 and relatively little in Scenario 2. Fig. 15.9 documents the results for our experiments in Zurich and Samara.

Subjects report that they are angry if the fourth group member contributes less than they did. This effect is certainly more pronounced in the scenario where they contributed 16 than in the scenario where they contributed 5. The difference is highly significant, both in the Zurich and the Samara sample ($p < 0.001$, Mann-Whitney tests). When the fourth group member contributes more than the pivotal subject, then people report to be quite happy. Surprisingly, subjects are equally happy about the contribution of 20 of the fourth member both when they have contributed 5 or 16 tokens. In other words, the gain in happiness seems not to depend on the own contribution, whereas the intensity of the negative emotions strongly depends on the own contribution.

When we compare the Zurich subjects with the Samara subjects, we find qualitatively very similar results. Yet, a striking difference is that the Samara subjects reported significantly less intensive negative emotions towards the free rider (for an own contribution of 16) than the Zurich subjects. Likewise, for the Samarians, the reported positive emotions were also highly significantly less intense than for the Zurich subjects ($p < 0.0001$, Mann-Whitney tests). Thus, there seem to be strong cross-societal differences in the reported emotions.

Overall, the results suggest that free riding causes negative emotions. Moreover, the emotional pattern is consistent with the hypothesis that emotions trig-

ger punishment. First, the majority of punishments are executed by above-average contributors and imposed on below-average contributors. Second, recall that punishment increases with the deviation of the free rider from the other members' average contribution. This is consistent with the observation that negative emotions are the more intense the more the free rider deviates from the others' average contribution. Third, evidence from neuroscientific experiments supports the interpretation that emotions trigger punishment. For instance, Sanfey et al. (2003) had their subjects play the ultimatum game, while the subjects' brains were scanned (using fMRI). The ultimatum game (invented by Güth et al. 1982) is a two-player game in which player 1 is asked to split an amount of money, say 10 Euros, between him- or herself and a player 2. Player 2 can only accept or reject the proposal. If he accepts, the offer is implemented; if he rejects, both get nothing. A rejection of a positive offer in the ultimatum game is also an instance of altruistic punishment. The brain scans showed that in the recipients who received an unfairly low offer by a human player 1, areas in the brain lit up that are related to negative emotions. When the unfair offer came from a computerized player 1, recipients were much less negatively aroused. Bosman & van Winden (2002) investigated the 'power-to-take game', which is related to the ultimatum game. They elicited self-reported emotions and found that unfair behavior triggers negative emotions that are correlated with punishment. de Quervain et al. (2004) studied neural activations of punishing subjects. They found that punishment activates the 'reward centre' of the brain; i.e., to punish is rewarding. Hence, the proverb "revenge is sweet". They were also able to show that subjects, for whom punishment was more rewarding, actually punished more. Taken together, these regularities are consistent with the view that emotions are an important factor in the process triggering altruistic punishment. Yet, more research is certainly needed here. The emerging field of neuroeconomics (see Camerer et al., in prep.) will certainly play an important role in this endeavor.

15.6 The evolution of strong reciprocity

The evidence presented above shows that many people, but not all, behave reciprocally. They reward nice behavior and punish misdeeds. Since this takes place even in one-shot games, this kind of reciprocity has been termed 'strong reciprocity' (e.g. Gintis 2000), to distinguish it from reciprocal altruism that occurs in repeated games. Reciprocal altruism is strategic reciprocity that can also be exhibited by a completely selfish individual, who would never cooperate or punish in a one-shot context. In economics, the kind of evidence presented in this chapter helped to pave the way for replacing the once ubiquitous selfishness assumption with more realistic assumptions about human's social preferences. A recent and very fruitful development in economic theory has been to take up the experimental evidence and model it. For instance, Fehr & Schmidt (1999) and Bolton & Ockenfels (2000) assume that people have a dislike for inequality. A free rider puts himself into a payoff advantage and inequality-averse people punish to reduce this inequality. Rabin (1993), Falk & Fischbacher (in prep.) and

Dufwenberg & Kirchsteiger (2004) assume that many people punish unkind intentions (to free ride reveals a greedy intention) and that they reward kind behavior (i.e. they cooperate to reward others' cooperation). Falk et al. (2004) show that intentions indeed play an important role in punishment since people also punish when they cannot diminish payoff inequities through punishment.

These new models, whose power extends beyond cooperation games, can be seen as proximate theories, but what explains the existence of strong reciprocity? Specifically, if sufficiently many people punish free riders sufficiently strongly, then free riders have no incentive to free ride anymore. Yet, why should anyone punish and not free ride on other's punishment, since altruistic punishment is just a second-order public good? The answer will probably be found in the evolutionary conditions of the human species that caused a propensity for strongly reciprocal behavior among a significant fraction of the population. The evidence presented suggests that strong reciprocity cannot easily be explained by kin selection (Hamilton 1964), reciprocal altruism (Trivers 1971, Axelrod & Hamilton 1981), indirect reciprocity (Alexander 1987, Nowak & Sigmund 1998) and by costly signaling theory (Zahavi & Zahavi 1997, Gintis et al. 2001).

In our view, one promising approach is 'gene-culture co-evolution' (Gintis 2000, Henrich & Boyd 2001, Bowles et al. 2003, Boyd et al. 2003, Gintis et al. 2003, Boyd & Richerson 2004). One line of reasoning (e.g. Boyd et al. 2003) goes as follows. Assume that in a population there are two behavioral types, cooperators and defectors. The cooperators incur a cost c to produce a benefit b that accrues to all group members. Defection is costless and produces no benefit. If the fraction of cooperators is x, then the expected payoff for cooperators is $bx - c$, whereas defectors get bx. Thus, the payoff difference is c, independent of the number of cooperators. Cooperators would always be at an evolutionary disadvantage under such circumstances. Now assume that there is a fraction y of 'punishers' who cooperate and punish defectors. Punishment reduces the payoff of the punished defector (by p) but also of the punishing subject (by k). The payoff of cooperators who cooperate but do not punish ('second-order free riders') is $b(x + y) - c$; the punished defectors get $b(x + y) - py$, and the punishers earn $b(x + y) - c - k(1 - x - y)$. If the cost of punishments exceed the costs of cooperation (i.e. if $py > c$), then cooperators have a higher fitness than defectors and the fitness disadvantage of punishers relative to the second-order free riders is $k(1 - x - y)$. Thus, punishment is altruistic and the cooperation and punishment game can have multiple equilibria.

This line of reasoning reveals two things. First, there is an important asymmetry between altruistic cooperation and punishment. In an environment without punishment, cooperators are always worse off than defectors, irrespective of how numerous they are. Second, by contrast to the first observation, the cost disadvantage of altruistic punishment declines as defection becomes infrequent because punishment is not needed anymore. The selection pressure against altruistic punishers is weak in this situation.

This latter observation suggests that within-group forces, like copying successful and frequent behavior (see Henrich & Boyd 2001) can stabilize cooperation. Boyd et al. (2003) formally investigate another mechanism, cultural group selection. Recall that in the presence of strong reciprocators the cooperation

game may have multiple equilibria, equilibria which imply cooperation, and defection equilibria. Different groups may settle at different equilibria. Here, cultural group selection may come into play. The main idea is that groups with more cooperators are more likely to win inter-group conflicts and are less likely to become extinct, because they may better survive during famine, manage their common resources better etc. (see also Soltis et al. 1995). Therefore, this kind of group selection will tend to increase cooperation because groups who arrived at a cooperative equilibrium are more likely to survive. Moreover, cooperative groups will tend to have more punishers. Since the within-group selection effect is weak if there is a lot of cooperation, cultural group selection can support the evolution of altruistic punishment and maintain it, once it is common. To test this intuition rigorously, Boyd et al. (2003) developed a simple model and simulated it for important parameters, like group size, migration rates between groups and the cost of being punished. The parameters were chosen to mimic likely evolutionary conditions. The simulation results are very interesting because they show that cultural group selection can support altruistic punishment under a wide range of parameters. First, in the absence of punishment, group selection can only sustain cooperation in very small groups, whereas in the presence of punishment, high and stable cooperation rates can be achieved even in large groups. Second, higher migration rates between groups decrease cooperation rates. If the cost of being punished is small, then cooperation breaks down. This result is also consistent with the experimental evidence (see Anderson & Putterman, in prep. and Carpenter 2004). The significance of this and related models is to show that individual selection and cultural factors, like conformism and group selection may coexist (and not be incompatible as in purely gene-based models) and can explain why strong reciprocity may survive. Of course, further models that highlight the links between individual and cultural group selection should and will arise.

We conclude this section with a short discussion of frequent critiques that are leveled at evolutionary explanations of strong reciprocity (see Johnson et al. 2003, Fehr & Gächter 2003 and Fehr & Henrich 2003). One critique concerns group selection. According to the critics, strong reciprocity is merely a byproduct of reciprocal altruism, indirect reciprocity, or signaling. The skepticism against group selection arguments is probably founded in the view that genetic group selection is an implausible mechanism (see also Sober & Wilson 1998). Yet, as the above account of the Boyd et al. (2003) model should make clear, cultural group selection models work completely differently from genetic group selection models.

The second line of critique is that strong reciprocity is a 'mal-adaptation' (see e.g. Johnson et al. 2003). According to this argument, humans evolved in small and mostly stable groups and thereby acquired the psychology needed for sustaining cooperation. Thus, the human brain applies ancient cooperative heuristics even in modern environments, where they are mal-adaptive. Humans did not evolve to play one-shot games and therefore, when they are in a novel environment like a one-shot game in the experimental lab, they behave as if they were in a repeated game. In our view, this argument is problematic for two reasons. First, it is obvious that people did not evolve to play one-shot lab experiments and the

strong reciprocity observed there does not represent adaptive behavior. Yet, lab experiments allow us to test to what extent people distinguish between one-shot and repeated games and to what extent they think strategically. As demonstrated repeatedly above (compare Figs. 15.1, 15.2 and 15.5, and the references therein), people cooperate substantially more with 'Partners' than with 'Strangers'. People also report stronger negative emotions when they are cheated by a 'Partner' than by a 'Stranger' (Fehr & Henrich 2003). Moreover, there is also systematic evidence that people respond strongly to increased costs of punishment; they punish less and therefore cooperate less (Fehr et al. 1997, Anderson & Putterman, in prep., Carpenter 2004). Second, research by anthropologists shows that group dispersal, migration and thereby the possibility of meeting strangers was quite common (see Fehr & Henrich 2003, in particular p. 69-76). Thus, vigilant individuals who are able to distinguish whether they deal with a 'Partner' or a 'Stranger' should have a fitness advantage.

Irrespective of one's take in this debate, one should notice that the phenomenon of strongly reciprocal behavior sheds new light on important economic issues (see Fehr & Gächter 2000a, Fehr et al. 2002 and Fehr & Fischbacher 2002). Even if strong reciprocity is a mal-adaptation, it is an important element in explaining patterns of human behavior.

15.7 Summary and conclusions

Humans have achieved a level of cooperation in large groups of genetically unrelated individuals that is outstanding in the animal kingdom. Understanding why this is so is a challenge for all social and behavioral sciences. A theoretically important question in all behavioral sciences is to establish to what extent the observed behavior can be explained by selfishness alone. People might cooperate for various (selfish) reasons. They might cooperate strategically to secure long-term benefits, to gain a favorable reputation in other social activities, to avoid social disapproval and punishment and to gain a high social status and approval. In reality, these motives are in most cases inextricably intertwined. In this paper, we have demonstrated that the experimental laboratory allows the researcher to separate motivations. The most important findings from experimental research are as follows:

- People cooperate even in one-shot PDs and public goods experiments.
- Relative to one-shot encounters, cooperation is strongly increased in stable groups.
- In the absence of communication and/or punishment, cooperation in randomly-composed groups is very fragile. Even stable groups cannot maintain cooperation.
- There seem to be two main types of players: (i) selfish free riders, who in one-shot experiments do not contribute to the public good but may cooperate strategically in repeated games and (ii) conditional cooperators who cooperate if others cooperate. In randomly-composed groups, the interaction of these two types of players explains why cooperation is fragile. The exception

to this rule is groups that are composed of like-minded cooperators, who know that the other group members share their cooperative attitude.
- Communication, possibilities for exchanging social (dis-)approval and reputation building substantially enhance cooperation. Yet, cooperation may still be fragile.
- Many people are prepared to punish free riders if they have the possibility to do so. Such punishment is often 'altruistic' because it can be observed even in one-shot games where the punishing subject does not benefit from induced cooperation. Altruistic punishment can substantially increase and stabilize cooperation.
- Negative emotions toward free riders may be a proximate mechanism that can explain altruistic punishment.

From a theoretical point of view, the most important observation is the existence of 'strong reciprocity', the fact that people are prepared to cooperate and to punish free riders even in anonymous one-shot encounters where there are no future interactions. While the existence of strong reciprocity can be considered an undisputed fact, evolutionary explanations are still open to debate.

Acknowledgments

We gratefully acknowledge financial support by the Grundlagenforschungsfonds of the University of St. Gallen through the research project "Soziale Interaktionen, Unternehmenskultur und Anreizgestaltung". We also thank Peter Kappeler, Carel van Schaik, and the participants of the Freilandtage in Göttingen 2003 for their very helpful comments.

References

Alexander, R. (1987). *The biology of moral systems.* New York: Basic Books.

Anderson, C. M. and L. Putterman (forthcoming). Do non-strategic sanctions obey the law of demand? *Games and Economic Behavior*.

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

Andreoni, J., and R. Croson (1998). Partners versus Strangers: Random Rematching in Public Goods Experiments, forthcoming in *Handbook of Experimental Economic Results*, ed. by C. Plott and V. Smith.

Andreoni, J., and J. Miller (1993). Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence. *Economic Journal* 103, 570-585.

Axelrod, R., W. D. Hamilton (1981). The Evolution of Cooperation. *Science* 211, 1390-1396.

Blau, P. (1964). *Exchange and Power in Social Life.* New Brunswick: Transaction Publishers.

Bochet, O., T. Page and L. Putterman (forthcoming). Communication and Punishment in Voluntary Contribution Experiments. *Journal of Economic Behavior and Organization.*

Bolton, G. and A. Ockenfels (2000). A Theory of Equity, Reciprocity and Competition, *American Economic Review* 100(1), 166-193.

Bosman, R. and F. van Winden (2002). Emotional Hazard in a Power-to-Take Experiment. *Economic Journal* 112(1), 147-169.

Bowles, S., J. Carpenter and H. Gintis (2001). Mutual Monitoring in Teams: Theory and Evidence on the Importance of Residual Claimancy and Reciprocity. Mimeo, University of Massachusetts, Amherst.

Bowles, S., E. Fehr and H. Gintis (2003). Strong reciprocity may evolve with and without group selection. *Theoretical Primatology Project Newsletter* 1(12), Supplement. Available online: http://www.robertwilliams.org/tpp/tpp_v1-12supp.html

Boyd, R., and P. J. Richerson (1988). The Evolution of Reciprocity in Sizable Groups, *Journal of Theoretical Biology* 132(3), 337-56.

Boyd, R., and P. J. Richerson (2004). *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago: University of Chicago Press.

Boyd, R., H. Gintis, S. Bowles, and P. J. Richerson (2003). Evolution of Altruistic Punishment," *Proceedings of the National Academy of Sciences* 100(6), 3531–3535.

Brosig, J., A. Ockenfels, and J. Weimann (2003). The Effect of Communication Media on Cooperation. *German Economic Review* 4(2), 217-241.

Camerer, C. F. (2003). *Behavioral Game Theory*. Princeton: Princeton University Press.

Camerer, C., G. Loewenstein and D. Prelec (forthcoming). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*.

Carpenter, J. (2004). The Demand for Punishment. Mimeo, Middlebury College.

Carpenter, J. (forthcoming). Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods. *Games and Economic Behavior*.

Carpenter, J., H.P. Matthews and O. Ong'ong'a (2004). Why Punish? Social Reciprocity and the Enforcement of Prosocial Norms. *Journal of Evolutionary Economics* 14(4), 407-430.

Clark, K., and M. Sefton (2001). The Sequential Prisoner's Dilemma: Evidence on Reciprocation. *Economic Journal* 111, 51-68.

Colman, A.M. (1999). *Game theory and its applications in the social and biological sciences.* London and New York: Routledge.

Cooper, R., D. DeJong, R. Forsythe and T. Ross (1996). Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games. *Games and Economic Behavior* 12, 187-318.

Dawes, R.M., J. McTavish, and H. Shaklee (1977). Behavior Communication and Assumptions about Other People's Behavior in a Commons Dilemma Situation. *Journal of Personality and Social Psychology* 35(1), 1-11.

Dawes, R.M. and A.J.C. van de Kragt and J.M. Orbell (1988). Not me or Thee, but We: The Importance of Group Identity in Eliciting Cooperation in Dilemma Situations - Experimental Manipulations. *Acta Psychologica* 68, 83-97.

De Quervain, D., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, E. Fehr (2004). The Neural Basis of Altruistic Punishment. *Science*, 305, 1254-1258.

Dufwenberg, M. and G. Kirchsteiger (2004). A Theory of Sequential Reciprocity, *Games and Economic Behavior* 47, 268-298.

Engelmann, D. and U. Fischbacher (2002). Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game. Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 132.

Falk, A. and U. Fischbacher (forthcoming). A Theory of Reciprocity. *Games and Economic Behavior*.

Falk, A., E. Fehr and U. Fischbacher (2004). Driving forces behind informal sanctions. IEW Working paper No. 59, University of Zurich.

Falk, A., S. Gächter, and J. Kovács (1999). Intrinsic Motivation and Extrinsic Incentives in a Repeated Game with Incomplete Contracts. *Journal of Economic Psychology* 20, 251-284.

Fehr, E. and U. Fischbacher (2002). Why social preferences matter – the impact of non-selfish motives on competition, cooperation and incentives. *Economic Journal* 112, C1-C33.

Fehr, E. and U. Fischbacher (2003). The Nature of Human Altruism. *Nature* 425, 785-791.

Fehr, E., U. Fischbacher, and S. Gächter (2002). Strong Reciprocity, Human Cooperation, and the Enforcement of Social Norms. *Human Nature* 13(1), 1-25.

Fehr, E., and S. Gächter (2000a). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives* 14(3), 159-181.

Fehr, E., and S. Gächter (2000b). Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90, 980-994.

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

Fehr, E. and S. Gächter (2003). The puzzle of human cooperation – reply. *Nature* 421, 912.

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

Fehr, E., and 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.* Cambridge: The MIT Press.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993). Does Fairness Prevent Market Clearing? *Quarterly Journal of Economics* 108, 437-459.

Fehr, E. and K. M. Schmidt (1999). A Theory of Fairness, Competition and Cooperation. *Quarterly Journal of Economics* 114(3), 817-868.

Fessler, D., and K.J. Haley (2003). The Strategy of Affect: Emotions in Human Cooperation. In: P. Hammerstein (ed.): *Genetic and Cultural Evolution of Cooperation.* Cambridge: The MIT Press.

Fischbacher, U. and S. Gächter (2004). Heterogeneous Social Preferences and the Dynamics of Free Riding in Public Goods. Mimeo, University of St. Gallen.

Fischbacher, U., S. Gächter, and E. Fehr (2001). Are People Conditionally Cooperative? Evidence from a Public Goods Experiment. *Economics Letters* 71, pp. 397-404.

Friedman, D. and S. Sunder (1994): *Experimental Methods. A Primer for Economists*. Princeton University Press.

Fudenberg D. and E. Maskin (1986). The Folk Theorem in Repeated Games with Discounting or with Incomplete Information, *Econometrica* 54, 533-556.

Gächter, S. and A. Falk (2002). Reputation and Reciprocity – Consequences for the Labour Relation. *Scandinavian Journal of Economics* 104(1), 1-26.

Gächter, S. and E. Fehr (1999). Collective Action as a Social Exchange. *Journal of Economic Behavior and Organization* 39, 341-369.

Gächter, S., and E. Renner (2004). Leading by Example in the Presence of Free Rider Incentives. Mimeo, University of Nottingham.

Gächter, S., and C. Thöni (forthcoming). Social Learning and Voluntary Cooperation among Like-Minded People. *Journal of the European Economic Association.*

Gächter, S., B. Herrmann and C. Thöni (2003). Understanding Determinants of Social Capital. Cooperation and Informal Sanctions in a Cross-Societal Perspective. Mimeo, University of St. Gallen.

Gintis, H. (2000). Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology* 206, 169-179.

Gintis, H., S. Bowles, R. Boyd, and E. Fehr (2003). Explaining Altruistic Behavior in Humans. *Evolution and Human Behavior* 24, 153-172.

Gintis, H., E. Smith and S. Bowles (2001). Costly Signaling and Cooperation, *Journal of Theoretical Biology* 213, 103-119.

Gürerk, Ö., B. Irlenbusch, and B. Rockenbach (2004). On the evolement of institutions in social dilemmas. Mimeo, University of Erfurt.

Güth, W., R. Schmittberger, and B. Schwarze (1982). An Experimental Analysis of Ultimatum Bargaining, *Journal of Economic Behavior and Organization* 3, 367-88.

Güth, W., M.V. Levati, E. van der Heijden and M. Sutter (2004). Leadership and Cooperation in Public Goods Experiments. Discussion Paper 28-2004, Papers on Strategic Interaction, Max Planck Institute for Research into Economic Systems.

Hamilton, W. D. (1964). Genetical Evolution of Social Behavior I, II. *Journal of Theoretical Biology* 7(1), 1-52.

Hammerstein, P. (ed). (2003a). *Genetic and Cultural Evolution of Cooperation*. Cambridge: MIT Press.

Hammerstein, P. (2003b). Why Is Reciprocity So Rare in Social Animals? A Protestant Appeal. In: P. Hammerstein (ed.): *Genetic and Cultural Evolution of Cooperation.* Cambridge: The MIT Press.

Hardin, G. (1968). The Tragedy of the Commons. *Science* 162, 1243-1248.

Henrich, J. and R. Boyd (2001). Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas. *Journal of Theoretical Biology* 208, 79–89.

Henrich J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis and R. McElreath (2001). In Search of Homo Economicus – Behavioral Experiments in 15 Small-Scale Societies, *American Economic Review* 91(2), 73-79.

Isaac, M., and J. Walker (1988). Communication and Free-Riding Behavior: The Voluntary Contribution Mechanism. *Economic Inquiry* 26(4), 585-608.

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

Kagel, John and Alvin E. Roth (eds) (1995). *Handbook of Experimental Economics.* Princeton: Princeton University Press

Keser, C. and F. van Winden (2000). Conditional Cooperation and Voluntary Contributions to Public Goods. *Scandinavian Journal of Economics* 102(1), 23-29.

Knauft, B. (1991). Violence and Sociality in Human Evolution. *Current Anthropology* 32(4), 391-428.

Kreps, D., P. Milgrom, J. Roberts, and R. Wilson (1982). Rational Cooperation in the Finitely Repeated Prisoners' Dilemma. *Journal of Economic Theory* 27, 245-252.

Ledyard, J. (1995). Public Goods: A Survey of Experimental Research. In Kagel, J. and A.E. Roth (eds): *Handbook of Experimental Economics*. Princeton: Princeton University Press.

Masclet, D., C. Noussair, S. Tucker, and M.-C. Villeval (2003). Monetary and Non-Monetary Punishment in the Voluntary Contributions Mechanism. *American Economic Review* 93(4), 366-380.

Milinski, M., D. Semmann and H.J. Krambeck (2002). Reputation Helps Solve the 'Tragedy of the Commons'. *Nature* 415, 424-426.

Moxnes, E. and E. van der Heijden (2003). The Effect of Leadership in a Public Bad Experiment. *Journal of Conflict Resolution* 47(6), 776-795.

Noussair, C., and S. Tucker (forthcoming). Combining Monetary and Social Sanctions to Promote Cooperation. *Economic Inquiry*.

Nowak M. and K. Sigmund (1998). Evolution of Indirect Reciprocity by Image Scoring. *Nature* 393, 573-577.

Oosterbeek, H., R. Sloof, and G. Van de Kuilen (2004). Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Experimental Economics* 7, 171-188.

Ostrom, E. J. Walker, and R. Gardner (1992). Covenants With and Without a Sword: Self-Governance is Possible. *American Political Science Review* 86(2), 404 – 417

Page, T., L. Putterman, and B. Unel (forthcoming). Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency. *Economic Journal*.

Palfrey, T.R., and J.E. Prisbrey (1997). Anomalous Behavior in Public Goods Experiments: How Much and Why? *American Economic Review* 87(5), 829–846.

Panchanathan, K. and R. Boyd (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* 432, 499-502.

Poundstone, W. (1992). *Prisoner's Dilemma*. New York: Anchor Books.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Review* 83(5), 1281-1302.

Rapoport, A., and A.M. Chammah (1965). *Prisoners' Dilemma. A Study in Conflict and Cooperation*. Ann Arbor: The University of Michigan Press.

Rege, M., and K. Telle (2004). The Impact of Social Approval and Framing on Cooperation in Public Good Situations. *Journal of Public Economics* 88, 1625-1644.

Richerson, P.J., R.T. Boyd, and J. Henrich (2003). Cultural Evolution of Human Cooperation. In: P. Hammerstein (ed.): *Genetic and Cultural Evolution of Cooperation.* Cambridge: The MIT Press.

Sally, D. (1995). Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992. *Rationality and Society* 7(1), 58-92.

Sanfey, A.G., J.K. Rilling, J.A. Aronson, L.E. Nystrom and J.D. Cohen, J.D. (2003). The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science* 300, 1755-1758.

Seabright, P. (2004): *The Company of Strangers. A Natural History of Economic Life.* Princeton: Princeton University Press.

Seinen, I. and A. Schram (forthcoming). Social Status and Group Norms: Indirect Reciprocity in a Helping Experimejnt. *European Economic Review*.

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

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

Wedekind, C. and M. Miliniski (2000). Cooperation Through Image Scoring in Humans. *Science* 288, 850-852.

Sefton, M., R. Shupp, and J. Walker (2002). The Effect of Rewards and Sanctions in Provision of Public Goods. CeDEx Working Paper 2002-2, University of Nottingham.

Selten, Reinhard and Rolf Stoecker (1986). End Behavior in Sequences of Finite Prisoner's Dilemma Supergames. A Learning Theory Approach. *Journal of Economic Behavior and Organization* 7, 47-70.

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

Yamagishi, T. (1986). The Provision of a Sanctioning System as a Public Good. *Journal of Personality and Social Psychology* 51(1), 110-116.

Zahavi, A. and A. Zahavi (1997). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. New York: Oxford University Press.