# Rationality and Commitment in Voluntary Cooperation: Insights from Experimental Economics

**Simon Gächter**
University of Nottingham
Centre for Decision Research and Experimental Economics (CeDEx)
School of Economics
Sir Clive Granger Building
University Park
Nottingham NG7 2RD
United Kingdom
simon.gaechter@nottingham.ac.uk

**Christian Thöni**
University of St. Gallen
Varnbuelstrasse 14
CH-9000 St. Gallen
Switzerland
christian.thoeni@unisg.ch

**Published in:** Fabienne Peter and Hans Bernhard Schmid (eds): ***Rationality and Commitment.*** Oxford: Oxford University Press 2007.

# 8

# Rationality and Commitment in Voluntary Cooperation: Insights from Experimental Economics

SIMON GÄCHTER AND CHRISTIAN THÖNI

## 1. The rationality of voluntary cooperation

Cooperation problems arise when individual incentives and social optimality diverge. This tension has intrigued social scientists and philosophers for decades.[1] In this chapter we look at the cooperation problem from the viewpoint of experimental economics, a subfield of economics which studies decision-making under controlled laboratory conditions and under real monetary incentives.[2] Years of careful experimentation have led to a body of results, which may shed new light on old philosophical questions, and on the foundations of the behavioural sciences. In particular our results will shed light on selfishness as one important foundational assumption of the behavioural sciences. The selfishness assumption has long been criticized (by Sen 1977, for instance, in a highly influential article). Yet only recently experimentalists have started to systematically scrutinize the selfishness assumption. We will discuss some selected evidence in this chapter on how people solve the cooperation problem and to what extent people's cooperation behaviour can be explained by their (non-)selfish preferences. We refer the reader to Fehr, Fischbacher and Gächter (2002), Camerer (2003), Fehr and Fischbacher (2003), and Hammerstein (2003) for broader discussions and surveys.

Our discussion will focus on two games, the Prisoner's Dilemma game and the public goods game. The widely known Prisoner's Dilemma game (PD from now on) is the prototype game in which this tension between individual and collective rationality arises. Figure 8.1 depicts a game between two players, which illustrates the issue.

| | | Column Player | |
|---|---|---|---|
| | | cooperate | defect |
| Row Player | cooperate | R ; R | S ; T |
| | defect | T ; S | P ; P |

Figure 8.1. The Prisoner's Dilemma game (if $T > R > P > S$)

When both players cooperate, both receive a payoff of R (for 'reward''). If the column player cooperates and the row player defects, then the row player receives the 'temptation payoff' T, and the cooperating column player the 'sucker's payoff' S. Payoffs are reversed if the row player cooperates and the column player defects. If both defect both receive the 'punishment payoff' P. This game is a PD, if $T > R > P > S$.[3]

It is now easy to see why the PD depicts a prototypical cooperation problem: if both cooperate both would be better off than when they both defect.[4] Yet, irrespective of the choice of the opponent each player always has a higher payoff (of either T or P) if he or she defects than if he or she cooperates. Thus, *if* the payoffs in the game of figure 8.1 obey $T > R > P > S$, *then* it is a PD. Rational players will therefore defect, since defection is a dominant strategy for both of them. They will defect even if they fully understand that mutual cooperation would collectively yield them a higher payoff than mutual defection, which is the only Nash equilibrium in this game.

The PD has intrigued researchers for decades (see Poundstone 1992 for an interesting discussion). It is probably one of the most extensively investigated games, both theoretically and experimentally. The empirical results from many experiments are equally as stark as the theoretical prediction of mutual defection. In a series of early experiments on the PD, Rapoport and Chammah (1965) found mutual cooperation in 30 to 50 per cent of all cases. This result has been replicated many times by now (see, e.g., Dawes 1980; Andreoni and Miller 1993; Ledyard 1995; Cooper

et al. 1996). Oberholzer, Waldfogel, and White (2003) report a particularly striking result. They analyse behaviour in a television show called *Friend or Foe*. In this show two subjects play a PD-like game for high stakes (between \$200 and \$16,400). If both players play 'friend' then they share the stake at hand equally. If one player plays 'foe' and the other plays 'friend' then the former receives the whole pie while the latter receives nothing. In case both players play 'foe' they both earn nothing. The participants of this television show choose 'friend' in slightly more than half of the cases.

Though very insightful, the PD models a two-person cooperation problem. Yet, in reality, many interesting cooperation problems involve many people. The public goods game is a suitable tool for studying cooperation in groups of more than two players. It can be seen as an *n*-person version of a PD. In the public goods game (PG from now on) a number of players form a group and each player is endowed with *e* tokens. The players decide simultaneously how many of their tokens ($g_i$) they want to contribute to the public good. The tokens not contributed count automatically as private income. All individual contributions in the group are summed up to $G = \sum g_i$. A player's payoff results as

$$\pi_i = e - g_i + aG \quad (1)$$

where the parameter *a* is the marginal per capita return (MPCR). This parameter measures the private return from contributing to the common good. For the game to be a PG game this parameter has to be within $1 > a > 1/n$. The first part of this condition ensures that the players have a dominant strategy to contribute nothing to the public good. The second part of the condition ensures that it is socially beneficial to contribute.

The solution of the PG is straightforward under the assumption that (1) represents the players' preferences. Contributing to the public good yields a return of *a*. This is less than what could be earned when keeping the tokens for oneself. Therefore, independent of the others' actions, each player has an incentive to choose the lowest possible contribution. However, since every member of the group profits from a player's contribution the social return is *na*, which is larger than unity. Therefore, joint payoff is maximized when all players contribute their full endowment.

Like the PD this game belongs to the most extensively studied games in experimental economics. Ledyard (1995) reviews the literature and reports that in a typical public goods game subjects contribute on average between

40 and 60 per cent of their endowment. When the experiment is repeated, contributions decrease over time to very low levels.

In summary, in many cases people manage to achieve mutual cooperation despite the fact that defection would have led to higher earnings for them individually. Thus, we have an empirical puzzle: there is much more cooperation than is compatible with the stark theoretical predictions of defection and freeriding. As a consequence, people overall are much better off than with the 'rational choice' of defection and freeriding. In the everyday sense, for many people voluntary cooperation rather than freeriding seems the 'rational' thing to do.

How can we explain this? We look at three different explanations that have been advanced: (i) cooperation in one-shot games is erroneous and maladaptive; (ii) people's reasoning may differ from the individualistic approach applied above; and (iii) the PD or public goods game does not adequately reflect people's true preferences. These possibilities have important conceptual consequences to which we will return in the final section. Our focus in the remainder of this chapter will be mainly on the last hypothesis. We will therefore only briefly sketch the first two explanations.

(i) According to the maladaptation hypothesis, one may argue that most games in real life are in fact repeated games. From the theory of repeated games it is well known that 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 and 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 and Hamilton 1981). Its essence is the idea that favours are reciprocated ('I'll scratch your back if you'll scratch mine') and that unhelpful behaviour is reciprocated by withholding future help. Thus, in indefinitely repeated games even selfish individuals have an incentive to cooperate.

Why then do people cooperate in one-shot games, where there is no future interaction? One explanation is just errors and confusion. A more

refined explanation in terms of errors is that people adopt behavioural rules that are beneficial in repeated cooperation games to the artificial one-shot game they are in (see, for instance, by Binmore 1994; 1998). A related argument, advanced by some evolutionary theorists, is the 'maladaptation hypothesis' (e.g., Johnson, Stopka and Knights 2003), according to which 'human brains apply ancient tendencies to cooperate that persist in newer environments, even if they are maladaptive (heuristic rules that violate expected utility often make sense for common tasks in our evolutionary history).' One problem with this argument is that in experiments people immediately change their behaviour in repeated games with the same opponent. For instance, in the ten-period repeated PDs of Andreoni and Miller (1993) and Cooper et al. (1996) cooperation rates were two times higher than in the ten one-shot games against different opponents. Keser and van Winden (2000) and Fehr and Gächter (2000) got similar results in the repeated vs. one-shot PG (see figure 8.5 below). We will come back to the maladaptation explanation in section 5.

(ii) The reasoning that has led to the theoretical prediction of mutual defection in the PD and full freeriding in the PG is based on the standard approach of rational choice analysis: given a decision-maker's preference over certain outcomes, each decision-maker chooses the outcome that maximizes his or her preference. Thus, the decision-maker looks at the problem from his or her own *individual* perspective. Yet people may reason differently in that they see themselves as being team members and therefore ask 'what should *we as a team* do?' People who apply the team perspective think about the actions they choose from the team perspective. People who apply 'team-directed reasoning' (Sugden 2000) will cooperate in the PD and the PG. Sugden (1993; 2000), and Bacharach (1999) have formalized this psychologically intuitive idea.[5] To our knowledge, there is no systematic evidence on team reasoning. We will sketch below some arguments used by experimental subjects that are consistent with team reasoning.

(iii) A third explanation for observed cooperation is that many people's preferences in the PD or PG are not adequately described by the payoffs depicted in figure 8.1 or equation (1). To appreciate this argument, one has to notice that game theory assumes that the payoff numbers in figure 8.1 reflect people's preference ordering of all possible strategy combinations. Recall that the game of figure 8.1 is only a PD if preferences obey

$T > R > P > S$. *If* this is the case, and if people apply individual instead of team-directed reasoning, then there is no way around the conclusion that rational people will defect in the PD and free ride in the PG (see Binmore 1994 for an extensive discussion of this issue).

If one neglects explanations (i) and (ii) for a moment, the stylized fact that many people cooperate in such simple games like the PD and the PG suggests that the actual utilities that people derive do not correspond to the specified payoffs. To see this, notice that all experiments require the specification of payoffs for the subjects. In virtually all experiments people receive monetary payments that induce the incentive structure that gives rise to a PD or PG, i.e., monetary payoffs obey $T > R > P > S$, or payoff function (1) in the PG. From a revealed preference approach, the observation of a cooperative choice may reflect that people actually *prefer* cooperation over defection. Yet, this implies that the utility of cooperation exceeds the utility of defection. Assume this is true for both players. Then the game of figure 8.1 actually is an 'assurance game'', where mutual cooperation is an equilibrium. In other words, the material incentives may not fully reflect people's preferences. Elements other than people's own material well-being might be relevant as well. For instance, people might have other-regarding preferences and simply care for the well-being of others. They might feel guilty if they do not cooperate or they may feel committed to reciprocate if they believe that others cooperate.

This line of reasoning is not without difficulty from a methodological point of view. Without further discipline, one can 'rationalize' any outcome by specifying the appropriate preferences. For this reason, theorists have resisted opening 'Pandora's box' by specifying preferences that rationalize outcomes. This argument is correct in our view in the absence of empirical tools to measure (or infer) people's preferences. We believe (and hope to demonstrate in this chapter) that the tools of experimental economics (and some further instruments like neuroscientific methods) may allow us to learn about the structure of people's motivations. This information may then guide theory building by putting empirically-disciplined structure on preference assumptions. We will return to this issue in section 6.

In the remainder of this chapter we will first focus on measuring motivations that might explain why people cooperate. Our purpose is twofold. We

demonstrate some methods how one can learn about people's motivations beyond the pecuniary payoffs they receive as a function of their choices. We will also show that all methods produce substantial evidence against the selfishness assumption frequently invoked by behavioural scientists, most notably by economics. We show that rather than being selfish many people are 'conditional cooperators' who are committed to cooperation if others cooperate as well. However, freeriders, who never contribute to the public good, exist as well. In other words, there is substantial heterogeneity in people's cooperative attitudes. We will present evidence on the consequences of such heterogeneous motivations in section 3. We will show that observed patterns of cooperation can be explained by preference heterogeneity but not easily by errors. We will then look at the role of emotions in cooperation in section 4. Emotions are interesting because it has been argued that they serve as a 'commitment device' (Hirshleifer 1987; Frank 1988). Specifically, freeriding may trigger feelings of anger in the cheated person, who may then be disposed to punish the freerider. If sufficiently many people are prepared to punish the freeriders, then freeriding may not pay off any more and induce even selfish people to cooperate. The evidence presented in sections 2 to 4 shows that many people apparently have unselfish preferences. This begs an explanation. In section 5 we will therefore sketch some recently advanced arguments by evolutionary theorists (Boyd et al. 2003) that provide some ultimate account for observed preferences. Section 6 provides some concluding remarks on possible methodological consequences of the findings presented in this chapter.

## 2. Measuring motivations

We will discuss some methods in this section on how to learn about people's motivation to (not) contribute to public goods. We will start by presenting qualitative evidence from verbal protocols and will then discuss tighter methods to infer motivations. We will demonstrate that all methods yield the same qualitative conclusions: A majority of people is non-selfishly motivated. In particular, they are prepared to cooperate if others cooperate. An important minority is best described as being selfish.

*2.1 Reasoning*

A natural first way to explore people's motivations is to ask them about their motives. Gächter and Fehr (1999) did PG experiments where the subjects had to explain their contribution decision.[6] In the following we discuss some of the answers the subjects gave when choosing their contribution in the first period of a repeated PG game. The subjects who chose to contribute their full endowment (of 20 tokens) provided the following statements:

A. 'By my decision I expect to motivate my team mates to high contributions.'
B. 'Trial to achieve ''safely'' the maximum. I try to convince the others.'
C. 'For maximal payoff, 20 to the group account.'
D. 'This way we earn the most as a group.'

Statements A and B speak for the notion of *rational cooperation*, i.e., the subjects try to encourage other subjects to contribute by providing a good example. On the other hand, statements C and D rather point in the direction of *team-directed reasoning*. Casual inspection of all answers favours the notion of *rational cooperation* since approximately two thirds of the cooperative subjects provide some sort of 'motivating others' argument. Subjects with intermediate contributions often say that they face a trade-off between securing their own income and motivating other subjects to contribute, for example:

E. 'No full risk. Signal disposition to contribute to the group.'
F. 'Do not want to put in everything before I know how the rest of the group will act.'

Subjects with low contributions are either on the very cautious side or plain freeriders.'

G. 'I don't invest that much because I don't know yet whether the others are pro-social or egoistic. If the others are egoistic, I have a loss.'
H. 'Most will contribute to the project. Maximal earnings for me.'

Statements in later periods naturally depend on the course of the game. We have seen in the previous section that contributions typically erode throughout the experiment. Three exemplary statements from later periods are the following:

I. 'The average contribution is high. I will try to keep the level of the group's contribution for a while and rip off thereafter.'
J. 'I will contribute the average of the others' contributions in the previous round.'
K. 'Enough is enough, from now on I will keep everything for myself. Everyone profited from my contributions, now I have to think about myself.'

Statement I is again a nice indication for *rational cooperation*. On the other hand, statements J and K hint to a type of behaviour that will be of special interest in the next section, namely *conditional cooperation*. In fact, most of the statements contain some sort of conditional cooperation argument. Contributing to the public good is clearly seen as desirable. Yet, for the majority of subjects, the reaction of other group members is crucial. Statements J and K can also be seen as supporting evidence for the notion of *inequality aversion* or *reciprocity*. Subjects obviously do not like their income to fall short of the others' incomes.

### *2.2 Eliciting beliefs about others' contributions*

While the verbal reasoning statements are insightful and suggestive of underlying motivations, the statements do not allow drawing tight conclusions about motivations. Specifically, we do not know what this subject expects others to contribute. A direct way around this is of course to simply ask the subjects about their belief about other group members' contributions. A subject who contributes nothing and expects a positive contribution of the other subjects might be seen as a freerider. If the contribution and the belief are positively correlated we would call the subject a conditional cooperator. If the contribution is high irrespective of the belief we would call the subject an unconditional cooperator or an altruist.

Croson (2007) was among the first to elicit beliefs and to correlate it with subjects' contribution behaviour. She found a very high and statistically significant correlation of beliefs and contributions: Subjects who expected others to contribute a lot were more likely to contribute high amounts than subjects who expected others to free ride.

Croson (2007) did not look at individual behaviour. Her observation is that on average people behave conditionally cooperatively in that their contributions and beliefs are positively correlated. Fischbacher and Gächter

(2006) also elicited beliefs and replicated Croson's findings of a positive correlation between beliefs and contributions. At the individual level they find subjects who do show a positive correlation between beliefs and contributions, whereas other subjects contribute zero even if they believe that others contribute positive amounts.

### *2.3 Inferring preferences from sequential decisions*

The beliefs data reported above provide the first systematic account of heterogeneous motivations. However, beliefs are not fully conclusive for inferring underlying motivations. Here is why. Consider we observe a subject contributing zero in the PG game. The subject might be classified as egoistic. Yet the subject might also be a conditional cooperator with pessimistic beliefs about the other group members' contributions.

The problem of the belief dependency of a conditional cooperator's contribution decision can be solved by a very simple trick. Instead of letting the subjects decide simultaneously one can conduct the PD or PG game *sequentially*. In such a game one can observe the decisions of players who *know* the contributions of the other team members rather than just have a belief about it. Fehr, Kosfeld and Weibull (2003) conducted the PD game as shown in figure 8.1 in the sequential mode. Subjects received monetary payoffs that ensured that the incentive structure induced a PD. But, as explained above, monetary incentives might not coincide with preferences. To elicit actual preferences, Fehr et al. applied the following procedure: the row players have to indicate whether they cooperate or defect for both cases where the first-moving column player has cooperated and where he or she has defected. Fehr et al. now take a 'revealed preference approach', i.e., an individual's preference is derived from observed choices. To see this, notice that if a player chooses to cooperate, when he or she could have also chosen to defect, then she apparently has a preference for cooperation. Since there are four possible outcomes, four possible preference types can be inferred: (1) a row player who chooses defect for both choices of the first mover, is a selfish freerider; (2) a row player who chooses 'freeride' in case the column player chooses 'freeride' and contributes if the column player does so too reveals to be a reciprocating conditional cooperator; (3) a row player who contributes in any case can be classified as an altruistic unconditional cooperator and, finally, a row player who does the opposite of the column player (i.e., behaves anti-reciprocally) is called 'other'.[7]

The data reported by Fehr et al. (2003) show that the first two types are clearly the most important; 47 per cent of the subjects act like freeriders and 38 per cent show the pattern of conditional cooperators.[8] Unconditional cooperators make up about 9 per cent of the population and the remaining 6 per cent of the subjects prefer to choose the opposite action.

The elicited preferences can now be used to answer the question which game the players really are playing. To appreciate this question, recall that the game in figure 8.1 is only a PD if both players in the game of figure 8.1 have preferences such that for both of them $T > R$ and $P > S$, i.e., both are selfish. In all other cases, the game they really play is not a PD. Thus, if types would randomly and independently be matched to play the PD, then they would play a PD in 22 per cent of the matchings, given the results of Fehr et al. (2003).

### *2.4 Eliciting 'contribution functions'*

Fischbacher, Gächter and Fehr (2001) and Fischbacher and Gächter (2006) use a similar revealed preference method to infer people's contribution preferences in a PG as a function of other group members' contributions. Therefore, the subjects in their experiment do not choose one contribution but a contribution as a function of other group members' average contribution. The PG game is played in groups of four subjects and the payoff function is again the same as in (1). The game is played just once to avoid confounds with strategic considerations. Every subject has to indicate a contribution *conditional on the average others' contribution*, i.e. for each of the 21 possible values of the average others' contribution subjects have to enter the number of points they want to contribute.

Fischbacher et al. classify their subjects according to their contribution function. A subject is called a freerider if and only if he or she contributes zero in all 21 cases. A subject is called a conditional cooperator if the contribution schedule is a clearly positive function of the others' average contribution. A somewhat peculiar type is the triangle contributor whose contribution is increasing in the others' contributions for low values and decreasing for high others' contributions. Figure 8.2 illustrates the (average) contribution function of the different types in the experiments by Fischbacher and Gächter (2006).More than half of all subjects are conditional cooperators and 23 per cent are freeriders. The rest are either triangle contributors, or 'others'. Fischbacher et al. (2001), and Herrmann

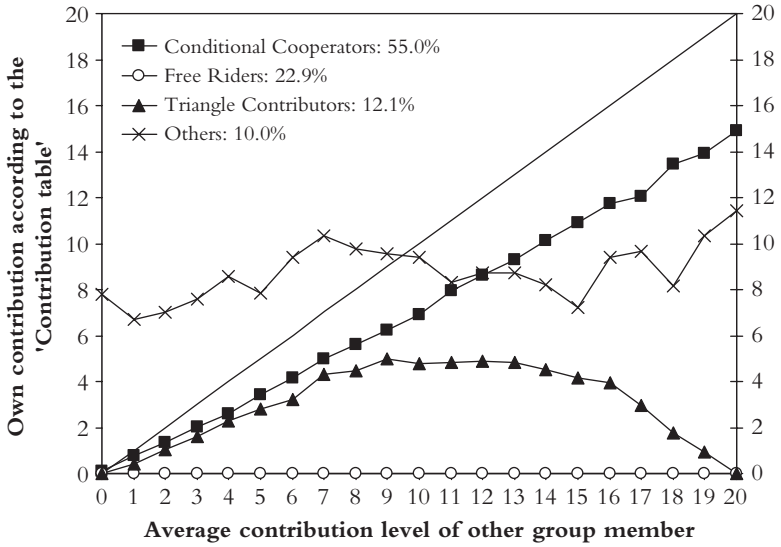


Figure 8.2. Average contribution function of types freerider, conditional cooperator, triangle contributor, and 'others'. Observations on the diagonal would correspond to the type of a perfect (i.e. one-to-one) conditional cooperator. Source: Fischbacher and Gächter (2006).

and Thöni (2007), who replicated this experiment in Russia, got very similar distribution of types and even of average contribution patterns.

### *2.5 Further methods*

There are further studies that try to understand preference heterogeneity. They use versions and/or combinations of the methods described above. Bardsley and Moffatt (2007), for instance, study sequential decisions in the PG game. The game is played in groups of seven people who face a similar payoff function as (1). Unlike the typical PG experiment, the game is played *sequentially*, i.e., the subjects choose their contributions consecutively. The authors are in particular interested in the way a subject's contribution decision is affected by the contributions of subjects deciding earlier. For a subject choosing early in the row there are several motives to choose a high contribution. On the one hand the subject might simply be cooperative. On the other hand, even a freerider might find it worthwhile to choose a high contribution if this induces other, later deciding subjects to contribute more. In other words, if a subject believes

that the subsequent subjects are conditionally cooperative then there is a strategic reason to choose a high contribution. Bardsley and Moffatt use econometric techniques to classify their subjects and find that 25 per cent are conditional cooperators, 25 per cent are freeriders, and the remaining 50 per cent contribute strategically. Since the latter contribute *only* strategically the authors conclude that, in a one-shot situation, they should count as freeriders as well. Therefore Bardsley and Moffatt characterize one quarter of their subjects as conditional cooperative and the remaining subjects as freeriders.

Kurzban and Houser (2005) report results from a similar PG experiment where the subjects first choose their contribution. Then the subjects are given the chance to change their contribution in a circular way. At every step, one of the subjects learns the actual group average and has to reconsider the own contribution decision. Kurzban and Houser classify 20 per cent as freeriders, 63 per cent as conditional cooperators, and 13 per cent as unconditional cooperators.

Burlando and Guala (2005) combine four different methods to assess a subject's type.[9] They find 32 per cent freeriders, 35 per cent conditional cooperators, and 18 per cent unconditional cooperators. The remaining 15 per cent cannot be classified.

Finally, the subjects in Muller et al. (2005) play a two-stage public goods experiment (using a variant of the strategy method). Muller et al. (2005) classify 35 per cent as selfish subjects who give nothing in the second stage irrespective of the first stage contribution of the other players; 38 per cent are conditional cooperators who condition their second stage contribution positively on the first stage contribution of the other players.

### *2.6 Summary*

Table 8.1 summarizes the results of the studies discussed in sections 2.3 to 2.5. Comparing the results across studies reveals that the distribution of types varies considerably. Clearly, the distribution of types is sensitive to the experimental tool used and the classification scheme. Some studies do not mention the unconditional cooperator as a special type. However, it is encouraging that, when using the same methodology, the numbers hardly differ. This is obvious when comparing Fischbacher et al. (2001), Fischbacher and Gächter (2006) and Herrmann and Thöni (2007) (in the latter study at least the fraction of conditional cooperators is similar to

the other studies). In addition to that, in the PG game where degrees of cooperativeness are allowed (as opposed to the PD game) it seems that the fraction of pure freeriders is between one fourth and one third of the population (with the exception of the study by Bardsley and Moffatt). The fraction of conditional cooperators seems to be substantially larger.

While numbers differ between studies, the significance of the findings summarized in Table 8.1 is that there is considerable heterogeneity in subjects' cooperative motivations. The results from the systematic preference elicitation experiments are consistent with the findings from the verbal protocols and the belief elicitation methods. A particularly noteworthy observation from this synopsis is the fact that freerider types are not ubiquitous. Many subjects have preferences that commit them for conditional cooperation. From the perspective of revealed preference theory, therefore, the game subjects really play may not be the PD game, or the public goods game as induced by the material incentives. The game subjects actually

Table 8.1. Overview of the distribution of types in Prisoner's Dilemma games (Fehr et al. 2003) and Public Goods games (all other studies)

| | Fehr et al. (2003) | Bardsley and Moffatt (2007) | Kurzban and Houser (2005) | Fischbacher et al. (2001) | Fischbacher and Gächter (2006) | Herrmann and Thöni (2007) | Burlando and Guala(2005) | Muller et al. (2005) |
|---|---|---|---|---|---|---|---|---|
| *N* | 96 | 98 | 84 | 44 | 140 | 148 | 92 | 60 |
| *Freeriders* (%) | 47 | 75 | 20 | 30 | 23 | 7 | 32 | 35 |
| *Conditional Cooperators* (%) | 38 | 25 | 63 | 50 | 55 | 57 | 35 | 38 |
| *Unconditional Cooperators* (%) | 9 | – | 13 | 2 | 1 | 2 | 18 | 3 |
| *Others / Unclassified* (%) | 6 | – | 4 | 18 | 21 | 34 | 15 | 24 |

play may well have multiple equilibria, in which cooperation may be an equilibrium outcome.

In the following section we discuss some consequences of this preference heterogeneity for the prospects of successful cooperation. We will show that conditional cooperators, who know that others are like-minded cooperators as well, manage to maintain very high and largely stable cooperation rates. By contrast, when groups consist of a mixture of types, cooperation almost inevitably unravels.

## 3. The consequences of heterogeneous motivations

Two immediate testable consequences of preference heterogeneity are that (i) in groups where group members are randomly selected cooperation is bound to be fragile and (ii) in groups that are composed of 'like-minded types' (i.e., groups composed of either cooperators or freeriders) we should see starkly different cooperation patterns. The reason for the first conjecture is that because freeriding types will not contribute, conditional cooperators will withdraw their cooperation and therefore cooperation is bound to collapse. The rationale for conjecture (ii) is that conditional cooperators who know that the other group members are 'like-minded' cooperators as well should be able to cooperate at a higher level than if they must fear the freeriders in their group. Groups composed solely of freerider types should not cooperate at all. In the following we discuss these two implications of preference heterogeneity in turn.

### *3.1 The instability of voluntary cooperation*

We provide evidence in this section that heterogeneous motivations in randomly composed groups will lead to fragile cooperation. The reason is that freeriders presumably do not contribute to the public good while the conditional cooperators' contributions might be non-minimal, depending on their belief about other group members' contributions. Subjects learn the contributions of the other team members during the repeated interaction. The freeriders have no reason to react to that information. The conditional cooperators on the other hand will update their beliefs. Given that the average conditional cooperator does not fully match the others' contribution the reaction will most likely be a reduction of contributions. There is no

reason to expect that the remaining types (triangle contributors and 'others') will behave in a way that offsets the negative trend.

To rigorously test this argument, Fischbacher and Gächter (2006) combined the elicitation of contribution functions described above with a standard ten-period public goods game played in the *stranger* mode, i.e., in every period the groups of four are formed randomly out of all subjects in a session. As predicted, contributions actually fall over time (from initially 40 per cent to 10 per cent by the last period).

Is this decline really due to the interaction of heterogeneously motivated types? A first hint is that the types (as identified by their contribution schedules) really contribute differently. The conditional cooperators contribute on average 28 per cent of the endowment while the freerider's average contribution is only 12 per cent. Surprisingly, also the freeriders contribute in the repeated game. However, looking at individual data Fischbacher and Gächter report that 70 per cent of the freeriders never choose a contribution above zero during the ten periods. Therefore, the majority of the subjects classified as freeriders do indeed freeride all the time. Among the conditional cooperators this fraction of subjects who always chose the minimal contribution during the ten periods is much lower at 25 per cent.

A second and more stringent support for the conjecture comes from using the elicited contribution functions for predicting contributions. Recall that the strategies asked subjects to indicate how much they are prepared to contribute to the public good for all feasible average contribution levels of the other group members. In the standard ten-period public goods game Fischbacher and Gächter (2006) also elicited in each period each subject's belief about the other group members' contributions. Therefore, we can—given a stated belief about other group members' average contribution—predict what a subject should contribute to the public good if he or she would be perfectly consistent with his or her elicited contribution function. Figure 8.3 depicts the actual average contributions in the ten rounds of the public goods game and the predicted contributions as a result of stated beliefs and contribution schedules.

Although average predicted contributions are too low compared with actual contributions, we find that predicted contributions, which are derived from the contribution functions and the elicited beliefs, actually decline. Therefore, this result supports the argument that preference heterogeneity rather than solely learning and reduced errors leads to unstable

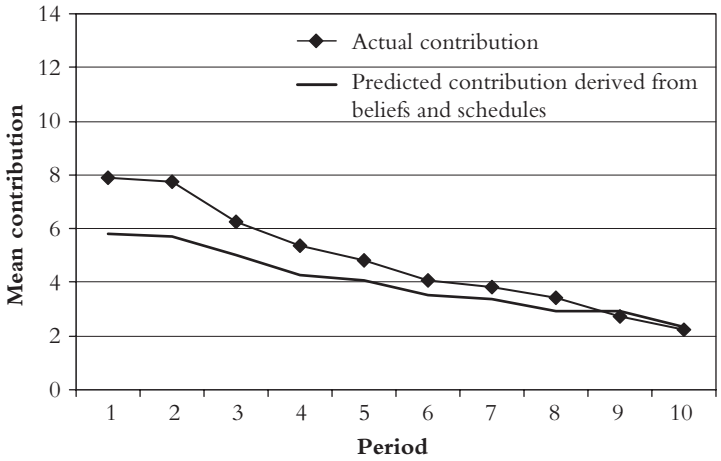


Figure 8.3. Average actual contributions and predicted contributions.
Source: Fischbacher and Gächter (2006).

cooperation. A further conceptually important implication of this result is that the interaction of heterogeneously motivated subjects may lead to freeriding *behaviour* despite the fact that not everyone is *motivated* by selfishness.

### *3.2. Voluntary cooperation among like-minded people*

We have seen that a mixture of conditional cooperators and freeriders is unfavourable for reaching cooperation in the PG game. According to our second conjecture, conditional cooperators would presumably prefer to play the game with like-minded cooperators. 'Team-directed reasoning' and subsequent cooperation should be easy if the team players know that they are among like-minded group members. Similarly, if the 'true game' subjects are playing is a game where cooperation is one of the equilibria (freeriding being another one), then knowing that others are like-minded cooperators should make it easy for subjects to coordinate on cooperation and to prevent freeriding. Likewise, if freerider types would know that they are among other freeriders, freeriding should be paramount.

Gächter and Thöni (2005) conducted an experiment where the subjects play in groups of 'like-minded' people. Thereby, like-mindedness refers to the type of a subject according to a classification whether one is a

freerider or a conditional cooperator. The experiment starts with a one-shot PG game. When all subjects have chosen their contribution the subjects are ranked according to their contribution. Then the subjects are reassigned to new groups. The reassignment works as follows. The three subjects with the highest contribution in the one-shot PG game constitute a first group. The subjects with the fourth to sixth highest contribution are in the second group and so on. Finally, the three least cooperative subjects find themselves in the last group. The subjects are informed about the reassignment procedure only after they finished the first game. Then the subjects learn the contributions their new group members chose in the one-shot PG game. In the new group subjects play a ten-period PG-game.

The reassignment mechanism adopted in this experiment sorts the subjects according to their contribution in a one-shot PG game. We have seen above that from the mere contribution decision we cannot determine the type of a subject unambiguously. This is because the type of conditional cooperator is compatible with all levels of contribution. We believe nevertheless that the reassignment mechanism as adopted by Gächter and Thöni provides a useful classification of the subjects along the dimension 'uncooperative–cooperative'. In addition, the mechanism is easy to understand from the subjects' point of view, which is crucial for the experiment. It is also important to note that the subjects do not know the reassignment mechanism when choosing their contribution in the one-shot PG game. A high contribution in this game therefore credibly reveals a cooperative attitude.

How do subjects play the PG game when they know they are among like-minded people? Gächter and Thöni (2005) report the results from eighteen groups of three subjects. The left panel of figure 8.4 shows the results of the main treatment. In this game the maximal contribution is 20. For expositional ease the groups are divided into three classes (TOP, MIDDLE and LOW) according to their average contribution in the one-shot PG game. The three graphs show the average contribution during the ten periods separated by class. The unconnected dots in period zero show the average contribution in the one-shot PG game that determines the group composition. The classes remain clearly separated over all periods. The groups in the TOP class consist to a large degree of subjects who contributed their entire endowment in the one-shot PG game. These

groups manage to maintain almost full cooperation until the penultimate period. The contributions of the MIDDLE class (consisting of subjects with intermediate contributions in the one-shot PG game) show a similar pattern on a somewhat lower level. Surprisingly, the subjects in the LOW class also, who almost all chose a contribution of zero in the one-shot PG game, manage to reach a certain level of cooperation in the repeated game.

The right panel of figure 8.4 shows the results from a control experiment. Groups are formed randomly as usual in this experiment, i.e., there is no reassignment according to cooperativeness. In order to make the two treatments comparable the data is still separated into the three classes. The separation now merely reflects the fact that there is variance in the contributions.

What does the comparison between the left and the right panel of figure 8.4 tell us about the effect of grouping like-minded subjects? First of all, cooperation in the TOP class of the sorted treatment is much higher than the average contribution in the random treatment (dotted line in the right panel). However, the real value of the sorting mechanism becomes clear if we compare the TOP class with the most cooperative third of the groups in the random treatment. The average contribution of the TOP class of like-minded groups is significantly higher than the average contribution of the most cooperative third of the groups in the random treatment. The fact that even the groups in the LOW class contribute somewhat more if they know they are among like-minded people is surprising at first sight. However, if uncooperative subjects know that they are among themselves then it is clear that there are no cooperative subjects to free ride on. This presumably motivates even uncooperative subjects to contribute some of their endowment in order to encourage the other freeriders to contribute as well. These groups might engage in 'rational cooperation' in the sense of Kreps et al. (1982). The fact that contributions drop to zero in the last period A supports this hypothesis.

These results are hard to reconcile with an error-hypothesis but are consistent with social learning (i.e., learning about the behaviour of others) by heterogeneous types. The reason is that an error hypothesis would not easily predict that group composition effects matter for cooperation behaviour. Since people are heterogeneous with respect to their attitudes to cooperation, the results suggest that the dynamics of cooperation as produced by social learning will depend very strongly on the extent to

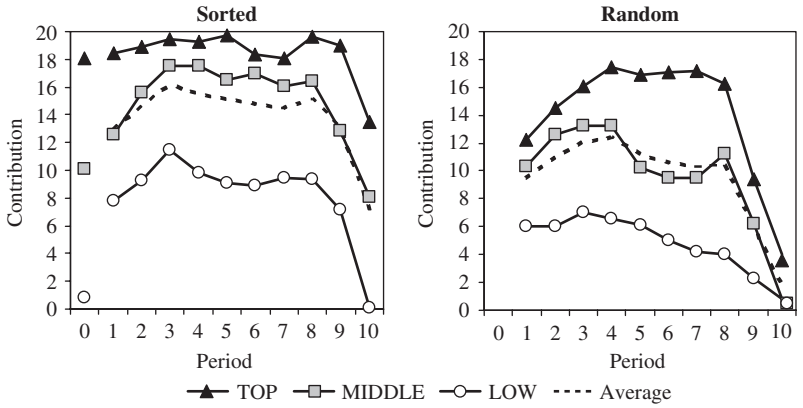


Figure 8.4. left panel: Average contributions over the ten periods for the TOP, MIDDLE and LOW class in the Sorted treatment. The unconnected dots in period zero are the average contributions in the Ranking treatment. Right panel: Average contribution of the most, intermediate and least cooperative groups over the ten periods.

Source: Gächter and Thöni (2005).

which group members are 'like-minded'. The results also confirm that social norms of cooperation are quite easy to sustain in homogeneous groups of people who are aware that others share their attitudes.

## 4. Altruistic punishment and negative emotions as a commitment device

The experiments discussed in the previous section have shown that the cooperation problem can be solved if the 'right' people are grouped together. However, the cooperation problem is thereby only solved for the groups consisting of very cooperative subjects. In mixed groups cooperation is bound to collapse, since conditional cooperators will reduce their contributions, once they realize that others free ride on them. Stopping cooperation is the only way to punish defectors. What if targeted punishment were possible?

In this section we will discuss a slight change in the PG game that allows for targeted punishment of group members. This game (as introduced by Fehr and Gächter 2000, 2002) has two stages. The first stage is identical to the usual PG game. In the second stage the subjects learn the contributions

of the other group members. They then have the possibility to punish each other by assigning 'negative points' to other group members. Each negative point costs one money unit to the punisher and reduces the income of the punished subject by three money units.

Why should such a mechanism change the behaviour in the PG game? According to standard economic theory (i.e., under the joint assumptions of rationality and selfishness) it would not. The reason for this lies in the fact that punishment is costly for both parties involved. Even if it is possible to 'educate' other team-members with the stick, the cooperative subjects of a group still have to solve a cooperation game. Punishing other subjects is itself a public good.

However, standard economic theory neglects a potentially influential factor, namely the subjects' emotions. Being 'suckered' is presumably a negative experience for most of us. Such negative emotions might trigger revenge. The PG game with punishment gives the subjects a much more precise measure to seek revenge than just to withhold cooperation. People can use the punishment option to eliminate the freerider's payoff advantage.

Panel A of figure 8.5 (adopted from Fehr and Gächter 2000) shows that the possibility of using informal sanctions indeed leads to significantly higher contributions relative to the PG without punishment opportunities. This is true for both repeated interactions (the so-called 'Partners'-treatment, where group composition stays constant and one-shot situations (the 'Strangers'-treatment, where group composition changes randomly from period to period). In the Partner-condition, contribution in the presence of punishment even approach almost full cooperation. Notice also that cooperation—both with and without punishment—is substantially higher among partners than among strangers. This holds already from the first period. We see this result as evidence against the maladaptation hypothesis discussed in section 1.

Panel B of figure 8.5 depicts the average punishment a subject has received for a given deviation of that subject's contribution from the average contribution of his or her group. The figure makes clear that more freeriding leads to more punishment. This holds for both partners and strangers. There is also no important difference in punishment between the two treatments, despite the fact that cooperation levels differ strongly. This suggests that the same deviation from a given group average is punished equally and punishment seems not to be used strategically.[10]

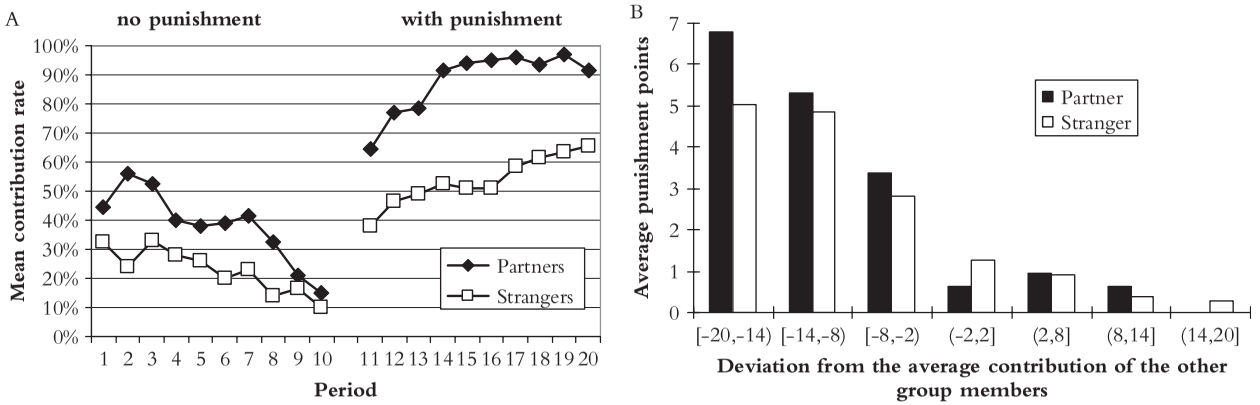


Figure 8.5. Cooperation patterns in the absence and presence of punishment and in stable ('Partner') and randomly changing ('Stranger') groups (panel A). Panel B: Mean received punishment as a function of one's deviation from the group average contribution.

Source: Fehr and Gächter (2000).

Subjects' cooperation behaviour looks quite rational given punishment. Subjects cooperate in accordance with the group norm to avoid punishment. Yet, why do people punish freeriders in a one-shot context although this is costly? Emotions may play a decisive role here (Fessler and Haley 2003) and negative emotions, in particular, may provide a *proximate* explanation. Freeriding may cause strong negative emotions among the cooperators and these emotions may trigger the willingness to punish the freeriders in turn. Theorists like Hirshleifer (1987) and Frank (1988) have argued that emotions may serve as a commitment device that induces people to retaliate if they feel cheated since an important property of emotions is that they imply an action tendency (see, e.g., Elster 1998).

If the conjecture is correct that freeriding triggers negative emotions, we should observe particular emotional patterns in response to freeriding. To elicit these patterns the participants of the Fehr and Gächter (2002) 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 freerider) on a seven point 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.

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. 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.

Overall, the results suggest that freeriding causes negative emotions. Moreover, the emotional pattern is consistent with the hypothesis that emotions trigger punishment. First, the majority of punishments are executed by above-average contributors and imposed on below-average contributors. Second, recall from figure 8.5B that punishment increases with the deviation of the freerider from the other members' average contribution. This is consistent with the observation that negative emotions are the more intense the more the freerider 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 is a two-player game in which player 1 is asked to split an amount of money, say €10, between him- or herself and a player 2. Player 2 can only accept or reject the proposal. The offer is implemented if he accepts; both get nothing if he rejects. A rejection of a positive offer in the ultimatum game is also an instance of 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. de Quervain et al. (2004) also 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 proximate mechanism in the process that triggers punishment. In the next section we look at ultimate explanations for cooperation and punishment.

## 5. Ultimate explanations

The evidence presented above shows that many people—but not all—behave reciprocally. They cooperate if others cooperate and they punish freeriders. Since this takes place even in one-shot games, this kind

of reciprocity has been termed 'strong reciprocity' (e.g., Gintis 2000; Fehr, Fischbacher and Gächter 2002, 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.

What explains the existence of strong reciprocity? Specifically, if sufficiently many people punish freeriders sufficiently strongly, freeriders 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 behaviour 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 and Hamilton 1981), indirect reciprocity (e.g., Nowak and Sigmund 1998) and by costly signalling theory (Zahavi and Zahavi 1997).

One explanation, already mentioned above is the 'maladaptation hypothesis''. According to this account, strong reciprocity is merely a by-product of reciprocal altruism, indirect reciprocity, or signalling. Humans evolved in small and mostly stable groups and thereby acquired the psychology and emotions needed for sustaining cooperation. Thus, the human brain applies ancient cooperative heuristics even in modern environments, where they are maladaptive. Humans did not evolve to play one-shot games and therefore, when they are in a novel environment like a one-shot lab experiment, they behave as if they were in a repeated game.

This argument is problematic in our view 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 behaviour. Yet, laboratory experiments allow us to test to what extent people distinguish between one-shot and repeated games and to what extent they think strategically. People cooperate substantially more with 'partners' than with 'strangers' as demonstrated above (see figure 8.5B). People also report stronger negative emotions when they are cheated by a 'partner' than by a 'stranger' (Fehr and Henrich 2003). Second, anthropologists have shown that group dispersal, migration and thereby the possibility to meet strangers was quite common (see Fehr and Henrich 2003, in particular pp. 69–76). Thus, vigilant individuals who are able to

distinguish whether they deal with a 'partner' or a 'stranger' should have a fitness advantage.

An alternative and, in our view, quite promising approach is 'gene-culture co-evolution' (e.g., Boyd, Gintis, Bowles and Richerson 2003; Boyd and Richerson 2004). One line of reasoning (e.g., Boyd et al. 2003) goes as follows. Assume that in a population there are two behavioural 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 don't punish (''second-order freeriders'') 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 freeriders 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 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 any more. The selection pressure against altruistic punishers is weak in this situation.

This latter observation suggests that within-group forces, such as copying successful and frequent behaviour, 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 survive during famine, manage their common resources better and so on. 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. Boyd et al. developed a simple model to test this intuition rigorously. They simulated the model 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 decreases cooperation rates. Cooperation breaks down if the cost of being punished is small.

The significance of this and related models is to show that individual selection and cultural factors, such as conformism and group selection, may coexist (and not be incompatible as in purely gene-based models) and can explain why strong reciprocity may survive and why we see the preferences we see.

## 6. Concluding discussion

In this essay we have demonstrated that carefully controlled laboratory experiments allow for a systematic investigation of how people actually solve cooperation problems that by definition involve a tension between collective and individual interests. We have also shown that according to several different experimental instruments apparently many people have non-selfish preferences that dispose them to cooperate if others cooperate as well. Under appropriate interaction structures (being among like-minded cooperators or having punishment opportunities) these preferences allow people to realize much higher gains from cooperation than would be

possible if all were selfish. However, all instruments have also shown that a non-negligible fraction of subjects behaves selfishly and free rides in one-shot games with no punishment.

We believe that these results have important methodological consequences. First, the experiments have demonstrated that one *can* collect empirical information on the structure of people's preferences. In other words, it is not necessary to treat preferences as a black box any more.

Second, the results challenge the ubiquity of the selfishness assumption that underlies many models in the social sciences, in particular in economics, and in biology. In addition to the material payoff, many people are apparently also motivated by other-regarding considerations, and/or by reciprocity. In fairness, one should notice that abstract economic theory does *not* invoke selfishness. Preferences only need to obey consistency axioms and can otherwise encompass any sort of motivation. Yet, in practice, selfishness is often invoked.

Third, while most experimental facts that we and others have interpreted as strong reciprocity are undisputed by now because they have been often replicated, the ultimate explanation of strong reciprocity is still open to debate. We believe that the issue of providing an ultimate account of strong reciprocity is of considerable theoretical interest for all behavioural scientists who hitherto have based their models on selfishness as being the only evolutionary reasonable assumption.

Fourth, the observed results do not contradict rational choice theory. They only contradict the universal selfishness assumption. The experimental evidence from many different games including those discussed here (see Camerer 2003) has led to the development of rational choice models of 'social preferences'', which take others' well-being into account or model their taste for reciprocity. These models are standard game-theoretic models that put structure on people's utility functions and otherwise apply standard solution concepts like subgame perfect Nash equilibrium. For instance, Fehr and Schmidt (1999) or Bolton and Ockenfels (2000) propose a utility function with *inequality aversion*. In addition to own material payoff, people experience some disutility both if there is disadvantageous or advantageous inequality in the payoff allocation. In the former, one receives a smaller payoff than the comparison partner; in the latter one earns more than the comparison partner. Inequality-averse subjects might choose non-minimal contribution in equilibrium if they believe that the other subjects do so as

well. This is because they do not like their income to be higher than the others' incomes, i.e., they dislike advantageous inequality aversion. Other models like Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006) propose *reciprocal preferences*. Reciprocally motivated subjects might choose a non-minimal contribution if they believe the other subjects to do so as well in order to return the favour.[11]

Fifth, one may criticize these models because people's reasoning may not be individualistic as typically assumed in the rational choice approach but may be 'team-directed' (Bacharach 1999; Sugden 2000). The casual evidence that we have sketched above suggests this to be a psychologically plausible explanation of how people reason in games (although the approaches by Bacharach and Sugden are more philosophical than psychological).

Sixth, a further criticism is that people's rationality may fall short of the high rationality required in this sort of rational choice models because people are boundedly rational (see, e.g., Brandstätter, Güth and Kliemt 2003 for a recent encompassing discussion). We agree to this criticism but notice that these models were first steps in demonstrating that the relaxation of the selfishness assumption has led to many important insights that appeared puzzling before the new models were available.

**Notes**

1. See, e.g., Binmore (1994, 1998) for an extensive treatment.
2. 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 earned considerable amounts of money depending on their decisions. Thus, the laboratory allows observing real economic behaviour under controlled circumstances (see Friedman and Sunder 1994 for an introduction to methods in experimental economics; Kagel and Roth 1995; Camerer 2003 for overviews of important results; and Guala 2005 for a thorough discussion of the methodology of experimental economics).
3. The original story goes as follows. Two arrested criminals are interrogated separately and have to decide whether to confess or not to confess. If

both criminals do not confess they have to stay in jail for a short time. If one confesses while the other does not then the confessor can leave the prison while the other stays in jail for a long time. If both criminals confess they both stay in jail for an intermediate time.

4. Since the PD (and an *n*-person version of the PD, the public goods game introduced below) highlight the tension between individual and collective rationality they have been used to analyse 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.
5. We regard team-direct reasoning as psychologically plausible because team-directed reasoning is consistent with the observation that group identity (''we-feelings'') are important for cooperation (see Dawes 1980 and Dawes, van de Kragt and Orbell 1988).
6. Gächter and Fehr (1999) did not report the statements for lack of space.
7. In general, utilities in games are von-Neumann-Morgenstern utilities. To infer them, one would have to elicit more than just ordinal rankings as Fehr et al. do. For an insightful discussion see Weibull (2004). Yet, for the purposes of solving the game with the concept of strict Nash equilibrium, ordinal preferences are sufficient.
8. Clark and Sefton (2001) also study a sequential PD and find that between 37 and 42 per cent of the subjects cooperate conditionally on others' cooperation.
9. The four methods are: (i) the strategy method, similar to one described in section 2.4; (ii) a value orientation test devised by psychologists; (iii) a repeated public goods game; and (iv) a post-experimental questionnaire.
10. By now, these results have been replicated many times. For a survey of the most important results see Kosfeld and Riedl (2004).
11. Sugden (1984) was one of the first to argue for the importance of reciprocity in the voluntary provision of public goods.

**References**

Andreoni, J., and Miller, J., 1993. 'Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence'. *Economic Journal* 103: 570–85.

Axelrod, R., and Hamilton, W. 1981. 'The Evolution of Cooperation'. *Science* 211: 1390–6.

Bacharach, M. 1999. 'Interactive Team Reasoning: A Contribution to the Theory of Co-operation'. *Research in Economics* 53: 117–47.

Bardsley, N., and Moffatt, G. 2007. 'The Experimetrics of Public Goods: Inferring Motivations from Contributions'. *Theory and Decision* 62: 161–93.

Binmore, K. 1994. *Game Theory and the Social Contract Vol. 1: Playing Fair.* Cambridge, MA: MIT Press.

Binmore, K. 1998. *Game Theory and the Social Contract Vol. 2: Just Playing.* Cambridge, MA: MIT Press.

Bolton, G., and Ockenfels, A. 2000. 'ERC: A Theory of Equity, Reciprocity, and Competition'. *American Economic Review*, 90/1: 166–93.

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

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. J. 2003. 'Evolution of Altruistic Punishment'. *Proceedings of the National Academy of Sciences* 100/6: 3531–5.

Brandstätter, H., Güth, W., and Kliemt, H. 2003. 'The Bounds of Rationality: Philosophical, Psychological and Economic Aspects of Choice Making'. *Homo Oeconomicus* 20/2-3: 303–56.

Burlando, R., and Guala, F. 2005. 'Heterogeneous Agents in Public Goods Experiments'. *Experimental Economics* 8/1: 35–54.

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

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

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

Croson, R. 2007. 'Theories of Commitment, Altruism and Reciprocity: Evidence from Linear Public Goods Games'. *Economic Inquiry* 45: 199–216.

Dawes, R. M., 1980. 'Social Dilemmas'. *Annual Review of Psychology* 31: 169–93.

Dawes, R. M., van de Kragt, A. J. C., and Orbell, J. M. 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, J.-F., Fischbacher, U., Treyer, Y., Schellhammer, M., Schynyder, U., Buck, A., and Fehr, E. 2004. 'The Neural Basis of Altruistic Punishment'. *Science* 305: 1254–8.

Dufwenberg, M., and Kirchsteiger, G. 2004. 'A Theory of Sequential Reciprocity'. *Games and Economic Behavior* 47: 268–98.

Elster, J. 1998. 'Emotions and Economic Theory'. *Journal of Economic Literature* 36: 47–74.

Falk, A., and Fischbacher, U. 2006. 'A Theory of Reciprocity'. *Games and Economic Behavior* 54: 293–315.

Fehr, E., and Fischbacher, U. 2003. 'The Nature of Human Altruism'. *Nature* 425: 785–91.

Fehr, E., and Gächter, S. 2000. 'Cooperation and Punishment in Public Goods Experiments'. *American Economic Review*, 90/4: 980–94.

Fehr, E., and Gächter, S. 2002. 'Altruistic Punishment in Humans'. *Nature* 415: 137–40.

Fehr, E., and Henrich, J. 2003. 'Is Strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism'. In P. Hammerstein (ed.), *Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: MIT Press, pp. 55–82.

Fehr, E., and Schmidt, K. 1999. 'A Theory of Fairness, Competition, and Cooperation'. *Quarterly Journal of Economics* 114: 817–68.

Fehr, E., Fischbacher, U., and Gächter, S. 2002. 'Strong Reciprocity, Human Cooperation, and the Enforcement of Social Norms'. *Human Nature* 13/1: 1–25.

Fehr, E., Kosfeld, M., and Weibull, J.W. 2003. 'The Game Prisoners (Really) Play'. *Mimeo*, Institute for Empirical Research in Economics, University of Zurich.

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

Fischbacher, U., and Gächter, S. 2006. Heterogeneous Motivations and the Dynamics of Freeriding in Public Goods. *CeDEx Discussion Paper No. 2006–1*, University of Nottingham.

Fischbacher, U., Gächter S. and Fehr E. 2001. 'Are People Conditionally Cooperative? Evidence from a Public Goods Experiment'. *Economics Letters* 71: 397–404.

Frank, R. 1988. *Passion Within Reason. The Strategic Role of the Emotions*. New York: W.W. Norton & Company.

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

Fudenberg, D. and Maskin, E. 1986. 'The Folk Theorem in Repeated Games with Discounting or with Incomplete Information'. *Econometrica* 54: 533–56.

Gächter, S. and Fehr, E. 1999. 'Collective Action as a Social Exchange'. *Journal of Economic Behavior and Organization* 39: 341–69.

Gächter, S. and Thöni, C. 2005. 'Social Learning and Voluntary Cooperation Among Like-minded People'. *Journal of the European Economic Association* 3/2-3: 303–14.

Gintis, H., 2000. 'Strong Reciprocity and Human Sociality'. *Journal of Theoretical Biology* 206: 169–79.

Guala, F., 2005. *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Hamilton, W. 1964. 'Genetical Evolution of Social Behavior I, II'. *Journal of Theoretical Biology* 7/1: 1–52.

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

Herrmann, B. and Thöni, C. 2007. 'Measuring Conditional Co-operation'. *Mimeo*, University of Nottingham.

Hirshleifer, J. 1987. 'On the Emotions as Guarantors of Threats and Promises'. In J. Dupré (ed.), *The Latest on the Best. Essays on Evolution and Optimality*. Cambridge, MA: MIT Press.

Johnson, D., Stopka, P. and Knights, S. 2003. 'The Puzzle of Human Cooperation'. *Nature* 421: 911–12.

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

Keser, C. and van Winden, F. 2000. 'Conditional Cooperation and Voluntary Contributions to Public Goods'. *Scandinavian Journal of Economics* 102/1: 23–9.

Kosfeld, M. and Riedl, A. 2004. 'The Design of (De)centralized Punishment Institutions for Sustaining Cooperation'. Tinbergen Institute Discussion Paper TI 2004-025/1.

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

Kurzban, R., and Houser, D. 2005. 'An Experimental Investigation of Cooperative Types in Human Groups: A Complement to Evolutionary Theory and Simulations'. *Proceedings of the National Academy of Sciences* 102/5: 1803–7.

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

Muller, L., Sefton, M., Steinberg, R., and Vesterlund, L. 2005. 'Strategic Behavior and Learning in Repeated Voluntary-Contribution Experiments'. CeDEx Working Paper No. 2005-13, University of Nottingham.

Nowak M. and Sigmund, K. 1998. 'Evolution of Indirect Reciprocity by Image Scoring'. *Nature* 393: 573–7.

Oberholzer-Gee, F., Waldfogel, J., and White, M. 2003. 'Social Learning and Coordination in High-Stakes Games: Evidence From Friend or Foe'. *NBER Working Paper* 9805.

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 Chammah, A. M. 1965. *Prisoner's Dilemma. A Study in Conflict and Cooperation*. Ann Arbor: The University of Michigan Press.

Sanfey, A. G., Rilling, J. K., Aronson J. A., Nystrom L. E., Cohen, J. D. 2003. 'The Neural Basis of Economic Decision-making in the Ultimatum Game'. *Science* 300: 1755–8.

Sen, A. 1977. 'Rational Fools: A Critique of the Behavioral Foundations of Economic Theory'. *Philosophy and Public Affairs* 6: 317–44.

Sugden, R. 1984. 'Reciprocity: The Supply of Public Goods Through Voluntary Contributions'. *Economic Journal* 94: 772–87.

Sugden, R. 1993. 'Thinking as a Team: Toward an Explanation of Nonselfish Behaviour'. *Social Philosophy and Policy* 10: 69–89.

Sugden, R. 2000. 'Team Preferences'. *Economics & Philosophy* 16: 175–204.

Trivers, R. L. 1971. 'The Evolution of Reciprocal Altruism'. *Quarterly Review of Biology* 46: 35–57.

Weibull, J. 2004. 'Testing Game Theory'. In: S. Huck (ed.), *Advances in Understanding Strategic Behavior. Game Theory, Experiments, and Bounded Rationality*. Houndmills: Palgrave Macmillan, pp. 85–104.

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