Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 55

Appropriating the Commons - A Theoretical Explanation

Armin Falk, Ernst Fehr and Urs Fischbacher

September 2000

Appropriating the Commons - A Theoretical Explanation

Armin Falk, Ernst Fehr and Urs Fischbacher

Institute for Empirical Research in Economics

University of Zurich

Blumlisalpstr. 10, CH-8006 Zurich

September 2000

Abstract. In this paper we show that a simple model of reciprocal preferences explains major experimental regularities of common pool resource (CPR) experiments. The evidence indicates that in standard CPR games without communication and without sanctioning possibilities inefficient excess appropriation is the rule. However, when communication or informal sanctions are available appropriation behavior is more efficient. Our analysis shows that these regularities arise naturally when a fraction of the subjects exhibits reciprocal preferences.

Keywords: common pool resources, experiments, fairness, reciprocity, game theory, fairness models

JEL: C7, C91, C92, D00, D63, H41

1. Introduction

In his classic account of social dilemma situations, Hardin (1968) develops his pessimistic view of the 'tragedy of the commons'. Given the incentive structure of social dilemmas he predicts inefficient excess appropriation of common pool resources (CPR). Hardin's view has been challenged by the insights of numerous field studies reported in the seminal book by Ostrom (1990). In this book the metaphor of a tragedy is replaced by the emphasis that people are able to govern the commons. It is shown that in many situations people are able to cooperate and improve their joint outcomes. Moreover, the reported field studies point towards the importance of behavioral factors, institutions and motivations. However, while it has been shown that these factors collectively influence behavior, it is of course, almost impossible to isolate the impact of individual factors.

This is why we need controlled laboratory experiments: Only in an experiment it is possible to study the role of each factor in isolation. In carefully varying the institutional environment the experimenter is able to disentangle the importance of different institutions and motivations. The regularities discovered in the lab can then be used to better understand the behavior in the field. In this paper we concentrate on three such empirical regularities which are reported in Walker, Gardner and Ostrom (1990), Ostrom, Walker and Gardner (1992)[1]. They first

[1]We also refer to the book by Ostrom, Gardner and Walker (1994) which summarizes and discusses the experimental findings. For an overview on experimental results see also Kopelman, Weber and Messick (this volume).

study a baseline situation which captures the central feature of all CPR problems: Due to negative externalities, individually rational decisions and socially optimal outcomes do not coincide. In a next step, the baseline treatment is enriched with two institutional features, the possibility of informal sanctions and the possibility to communicate. The empirical findings can be summarized as follows. In the baseline CPR experiment aggregate behavior is best described by the Nash-Equilibrium of selfish money maximizers. People excessively appropriate the CPR and thereby give rise to the 'tragedy' predicted by Hardin (1968). Giving subjects the possibility to sanction each other, however, strongly improves the prospects for cooperative behavior. This is surprising because sanctioning is costly and therefore *not* consistent with the assumptions that provide the basis for Hardin's pessimistic view, i.e., that subjects are selfish and rational. A similar observation holds for the communication environment. Allowing for communication also increases cooperative behavior. The resulting efficiency improvement is again inconsistent with the behavioral assumptions underlying Hardin's analysis because communication does not alter the material incentives.

Taken together we have therefore the following puzzle: In a sparse institutional environment people tend to overharvest common pool resources. In this sense the pessimistic predictions by Gordon (1954) and Hardin (1968) which are based on the assumptions of selfish preferences are supported. At the same time, however, we find the efficiency enhancing effect of informal sanctions and communication. This is in clear contradiction to the standard rational choice view because, why

should a rational and selfish individual sacrifice money in order to sanction the behavior of another subject? And why should a money maximizing subject reduce his or her appropriation level following some cheap talk? The question is more general: Why is the rational choice conception correct in one setting and wrong in another?

In this paper we suggest an integrated theoretical framework that is capable to explain this puzzle. We argue that the reported regularities are compatible with a model of human behavior that extends the standard rational choice approach and incorporates *preferences for reciprocity and equity*. The basic behavioral principle which is formalized in our model is that a substantial fraction of the subjects acts conditionally on what other subjects do. If others are nice or cooperative they act cooperatively as well, while if others are hostile they retaliate[2]. Our model also accounts for the fact that there are selfish subjects who behave in the way predicted by standard rational choice theory. We formally show that the interaction of these two diverse motivations (reciprocity and selfishness) and the institutional set-up is responsible for the observed experimental outcomes. In the absence of an institution which externally enforces efficient appropriation levels or which allows for informal sanctions, the selfish players are pivotal for the aggregate

[2]The importance of reciprocity has been established in dozens if not hundreds of experiments. For rewarding behavior in response to kind acts, see Fehr, Kirchsteiger and Riedl (1993) or Berg, Dickhaut and McCabe (1995). For punishing behavior in response to hostile acts, see Güth, Schmittberger and Schwarze (1982). Recent overviews are provided in Ostrom 1998 and Fehr and Gächter (2000b).

outcome. However, if there is an institutional set-up that enables people to impose informal sanctions or allows for communication, the reciprocal subjects discipline selfish players and thus shape the aggregate outcome. As a consequence there is less appropriation in CPR problems and higher voluntary contributions in public goods situations. Even though it is our main purpose in this paper to show that the approach is able to account for the seemingly contradictory evidence of CPR experiments, we think that the developed arguments are very general and likely to extend beyond the lab.

In the next section we briefly outline the basic structure of our approach and recently developed fairness models. In section 3 we apply our model to the standard CPR game and discuss the theoretical predictions in the light of the empirical findings. It also contains propositions for a CPR game with sanctioning opportunities as well as a discussion on the role of communication in the presence of reciprocal preferences. Section 4 contains a comparison of CPR results to those arrived in public goods games. Section 5 concludes.

2. Theoretical models of reciprocity and fairness

A large body of evidence indicates that fairness and reciprocity are powerful determinants of human behavior (for an overview see, e.g., Fehr and Gächter, 2000b). As a response to this evidence, various theories of reciprocity and fairness have

been developed.[3] These models assume that - in addition to their material self-interest - people also have a concern for fair treatments. The impressive feature of these models is that they are capable to correctly predict experimental outcomes in a wide variety of experimental games. Common to all of these models is the premise that the players' utility does not only depend on their own payoff but also on *the payoff(s) of the other player(s).* This assumption stands in sharp contrast to the standard economic model according to which subjects' utility is based solely on their own absolute payoff.

Besides this similarity there are also important differences between the fairness theories, however. These differences concern, e.g., (i) the relevant reference agent, (ii) the importance of fairness intentions attributed to other players' actions and (iii) whether people punish in order to reduce distributional inequity or in order to reduce the other players' payoffs. In Falk, Fehr and Fischbacher (2000) it is shown that in many relevant contexts the predictive power of the theories depends critically on how they model fairness motives. It turns out, for example, that subjects do not only sanction because they want to reduce distributional inequities but that the motive to punish unfair intentions is also a major determinant of sanctioning behavior. This means that approaches like Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) which model fairness concerns exclusively as

[3]The models are: Rabin (1993), Levine (1998), Bolton and Ockenfels (2000), Fehr and Schmidt (1999), Falk and Fischbacher (1999), Dufwenberg and Kirchsteiger (1999) and Charness and Rabin (2000).

a problem of fair payoff distributions are incomplete. It also means, however, that approaches which model reciprocal behavior solely as a response to fair or unfair intentions (Rabin (1993) and Dufwenberg and Kirchsteiger (1999)) are incomplete as well. The empirical evidence in Blount (1995) and Falk, Fehr and Fischbacher (2000) clearly indicates that approaches which rely on distributional concerns and on rewarding and sanctioning of intentions (as in Falk and Fischbacher (1999)) best capture the experimental regularities.

In the games analyzed in this paper most fairness theories yield similar predictions. Therefore we restrict our attention to the model of Fehr and Schmidt because it is relatively easy to apply in our context.

In the Fehr and Schmidt model, fairness is modeled as 'inequity aversion'. An individual is inequity averse if he dislikes outcomes that are perceived as inequitable. This definition raises, of course, the difficult question how individuals measure or perceive the fairness of outcomes. Fairness judgments are inevitably based on a kind of neutral reference outcome. The reference outcome that is used to evaluate a given situation is itself the product of complicated social comparison processes. In social psychology (Festinger, 1954; Homans, 1961; Adams, 1963) and sociology (Davis, 1959; Pollis, 1968; Runciman, 1966) the relevance of social comparison processes has been emphasized for a long time. One key insight of this literature is that relative material payoffs affect people's well being and behavior. As we will see below, without the assumption that at least for some people relative payoffs matter, it is difficult, if not impossible, to make sense of the

empirical regularities observed in CPR experiments. There is, moreover, direct empirical evidence suggesting the importance of relative payoffs. The results in Agell and Lundborg (1995) and Bewley (1998), for example, indicate that relative payoff considerations constitute an important constraint for the internal wage structure of firms. In addition, Clark and Oswald (1996) show that comparison incomes have a significant impact on overall job satisfaction. Strong evidence for the importance of relative payoffs is also provided by Loewenstein et al. (1989). These authors asked subjects to ordinally rank outcomes that differ in the distribution of payoffs between the subject and a comparison person. On the basis of these ordinal rankings the authors estimate how relative material payoffs enter the person's utility function. The results show that subjects exhibit a strong and robust aversion against disadvantageous inequality: For a given own income x_i subjects rank outcomes in which a comparison person earns more than x_i substantially lower than an outcome with equal material payoffs. Many subjects also exhibit an aversion against advantageous inequality although this effect seems to be significantly weaker than the aversion against disadvantageous inequality.

The determination of the relevant reference group and the relevant reference outcome for a given class of individuals is ultimately an empirical question. The social context, the saliency of particular agents and the social proximity among individuals are all likely to influence reference groups and outcomes[4]. Because in

[4]For a first attempt to endogenize the choice of reference agents or standards in a formal model see Falk and Knell (2000).

the following we restrict attention to individual behavior in economic experiments we have to make assumptions about reference groups and outcomes that are likely to prevail in this context. In the laboratory it is usually much simpler to define what is perceived as an equitable allocation by the subjects. The subjects enter the laboratory as equals, they don't know anything about each other, and they are allocated to different roles in the experiment at random. Thus, it is natural to assume that the reference group is simply the set of subjects playing against each other and that the reference point, i.e., the equitable outcome, is given by the egalitarian outcome.

So far we have stressed the importance of the concern for relative payoffs. This does not mean, however, that the absolute payoff should be viewed as a quantité negliable. Moreover, we do not claim that all people share a (similar) concern for an equitable share. In fact, many experiments have demonstrated the heterogeneity of subjects and the importance of absolute payoffs. Therefore, the Fehr and Schmidt model (as well as the other fairness models) allows for both, absolute and relative payoffs and the apparent heterogeneity of individuals.

More precisely, in the Fehr and Schmidt model it is assumed, that in addition to purely selfish subjects, there are subjects who dislike inequitable outcomes. They experience inequity if they are worse off in material terms than the other players in the experiment and they also feel inequity if they are better off. Moreover, it is assumed that, in general, subjects suffer more from inequity that is to their material disadvantage than from inequity that is to their material ad-

vantage. Formally, consider a set of n players indexed by $i \in \{1, ..., n\}$ and let $\pi = (\pi_1, ..., \pi_n)$ denote the vector of monetary payoffs. The utility function of player i is given by

$$U_i = \pi_i - \frac{\alpha_i}{n-1} \sum_{j, \pi_j > \pi_i} (\pi_j - \pi_i) - \frac{\beta_i}{n-1} \sum_{j, \pi_i > \pi_j} (\pi_i - \pi_j) \tag{2.1}$$

where

$$\alpha_i \geq \beta_i \geq 0 \text{ and } \beta_i < 1.$$

The first term in (1), π_i, is the material payoff of player i. The second term in (1) measures the utility loss from disadvantageous inequality, while the third term measures the loss from advantageous inequality. Figure 1 illustrates the utility of player i as a function of x_j for a given income x_i. Given his own monetary payoff x_i, player i's utility function obtains a maximum at $x_j = x_i$. The utility loss from disadvantageous inequality ($x_j > x_i$) is larger than the utility loss if player i is better off than player j ($x_j < x_i$).

Figure 1

To evaluate the implications of this utility function let us start with the two-player case. For simplicity the model assumes that the utility function is linear in inequality aversion as well as in x_i. Furthermore, the assumption $\alpha_i \geq \beta_i$ captures the idea that a player suffers more from inequality that is to his disadvantage. The above mentioned paper by Loewenstein et al. (1989) provides strong evidence that this assumption is, in general, valid. Note that $\alpha_i \geq \beta_i$ essentially means

that a subject is loss averse in social comparisons: Negative deviations from the reference outcome count more than positive deviations. The model also assumes that $0 \leq \beta_i < 1$. $\beta_i \geq 0$ means that the model rules out the existence of subjects who like to be better off than others. To interpret the restriction $\beta_i < 1$ suppose that player i has a higher monetary payoff than player j. In this case $\beta_i = 0.5$ implies that player i is just indifferent between keeping 1 Dollar to himself and giving this Dollar to player j. If $\beta_i = 1$, then player i is prepared to throw away 1 Dollar in order to reduce his advantage relative to player j which seems very implausible. This is why we do not consider the case $\beta_i \geq 1$. On the other hand, there is no justification to put an upper bound on α_i. To see this suppose that player i has a lower monetary payoff than player j. In this case player i is prepared to give up one Dollar of his own monetary payoff if this reduces the payoff of his opponent by $(1 + \alpha_i)/\alpha_i$ Dollars. For example, if $\alpha_i = 4$, then player i is willing to give up one Dollar if this reduces the payoff of his opponent by 1.25 Dollars.

If there are $n > 2$ players, player i compares his income to all other $n - 1$ players. In this case the disutility from inequality has been normalized by dividing the second and third term by $n - 1$. This normalization is necessary to make sure that the relative impact of inequality aversion on player i's total payoff is independent of the number of players. Furthermore, we assume for simplicity that the disutility from inequality is self-centered in the sense that player i compares himself to each of the other players, but he does not care per se about inequalities within the group of his opponents.

3. Theoretical Predictions

In the following we discuss the impact of inequity aversion in typical CPR games. The first game we analyze is a standard CPR game without communication and sanctioning opportunities. We proceed by analyzing games that add the possibilities of costly sanctioning and communication respectively. For all games we first derive the standard economic prediction, i.e., the Nash equilibrium assuming that everybody is selfish and rational. We contrast this prediction with experimental results and the prediction derived by our fairness model. Notice that in presenting the experimental results we restrict our attention to behavior of subjects in the final period since in the final period non-selfish behavior cannot be rationalized by the expectation of rewards in future periods. Furthermore, in the final period we have more confidence that the players fully understand the game that is being played.

3.1. The Standard CPR Game

In a standard CPR game each player is endowed with an endowment e. All n players in the group decide independently and simultaneously how much they want to appropriate from a CPR. Individual i's appropriation decision is denoted by x_i. The appropriation decision causes a cost c per unit of appropriation but also yields a revenue. While the cost is assumed to be independent of the decisions of the other group members, the revenue depends on the appropriation decisions of all players. More specifically the total revenue of all players from the CPR

is given by $f(\sum x_j)$ where $\sum x_j$ is the amount of total appropriation. For low levels of total appropriation $f(\sum x_j)$ is increasing in $\sum x_j$ but beyond a certain level $f(\sum x_j)$ is decreasing in $\sum x_j$. An individual subject i receives a fraction of $f(\sum x_j)$ according to the individual's share in total appropriation $\frac{x_i}{\sum x_j}$. Thus the total material payoff of i is given by:

$$\pi_i = e - cx_i + \left[\frac{x_i}{\sum x_j}\right] f(\sum x_j)$$

In the experiments of Walker, Gardner and Ostrom (1990) $e = 10$ (or 25) and $c = 5$. The total revenue is given by $f(\sum x_j) = 23 \sum x_j - .25(\sum x_j)^2$. Thus in this experiment material payoffs are:

$$\pi_i = 10 - 5x_i + \left[\frac{x_i}{\sum x_j}\right] [23 \sum x_j - .25(\sum x_j)^2]$$

Intuitively this is a social dilemma problem since individual i's appropriation decision x_i does not only affect player i's payoff but also that of all other players. Beyond a certain level of total appropriations, an increase in the appropriation of player i lowers the other players' revenue from the CPR. Since selfish players are concerned only with their own well-being they do not care about the negative externalities they impose on others. As we discuss below this leads to the typical inefficiencies that are characteristic for this type of dilemma games.

The above payoff function from Walker, Gordon and Ostrom can be transformed as $\pi_i = 10 + 18x_i - 0.25x_i \sum x_j$ or more abstractly as $\pi_i =: e + ax_i - bx_i \sum x_j$. As we will see this notation will be useful in the following discussion.

In this CPR game the standard economic prediction (assuming completely selfish and rational subjects) is as stated in Proposition 3.1[5]:

Proposition 3.1 (Selfish Nash Equilibrium). *If all players have purely selfish preferences, the unique Nash equilibrium is symmetric and individual appropriation levels are given by* $x_i^* = \frac{a}{b(n+1)}$.

In the following we denote this equilibrium as SNE (Selfish Nash Equilibrium) and the corresponding individual appropriation levels as x_{SNE}. As can be seen from Proposition 3.1, the individual contribution is independent of the endowment and it is decreasing in the number of players. In the specification of Ostrom, Gardner and Walker groups of eight players participated in the experiment. Thus in their experiment the predicted individual contribution amounts to $x_i^* = \frac{18}{0.25(8+1)} = 8$. Given the group size, total appropriation is 64. Compared to the social optimum of 36 this equilibrium yields substantial inefficiencies[6]. The point is that in their decisions, subjects ignore the negative externality imposed on the other players. Since players are assumed to care only about their own material payoff they simply don't care about such externalities.

How does the presence of inequity averse or reciprocally motivated subjects alter the standard economic prediction? In order to answer this question we will discuss two propositions. The first proposition considers symmetric equilibria whereas the second deals with asymmetric equilibria.

[5] All propositions are proved in the appendix.

[6] The derivation of the social optimum is given in the appendix.

It is useful to start our discussion of the properties of symmetric equilibria with the nature of the *best-response function* of an inequity averse subject. The best response function indicates the optimal appropriation response of an inequity averse player to the average appropriation of all other players. Figure 2 shows the best-response of an inequity averse subject (with positive α and β) and compares it to that of a selfish subject. The thin line represents the optimal appropriation of a selfish subject given the average appropriation level of the other group members.[7] As can be seen in Figure 2, a selfish player appropriates less the more the other group members appropriate. At the point where the best response function intersects the diagonal, the selfish Nash equilibrium (SNE) prevails. At this point the average appropriation level of the other $n-1$ players is 8 from the viewpoint of each individual player. Moreover, it is in the self-interest of each player to respond to this average appropriation of the $n-1$ other players with an own appropriation level of 8. Now look at the bold line in Figure 2. This line shows the best response behavior of an inequity averse subject. Four aspects of this function are important to emphasize.

First, in the area above the diagonal, i.e., where the other players appropriate *less* than in the SNE, the best response curve of an inequity averse player lies *below* that of a selfish player. This means that if the other players are 'nice' in the sense that they appropriate less than what is in their material self interest,

[7]Note that Figure 2 shows the symmetric case, where the other players' appropriation decisions are equal.

an inequity averse subject also appropriates less. Since inequity averse players dislike being in a too favorable position they do not exploit the kindness of the other players but instead voluntarily sacrifice some of their resources in favor of the other players.

Second, there is an area below the diagonal. In this area the other group members appropriate more than in the SNE. The best response behavior of an inequity averse subject dictates to appropriate *more* than is compatible with pure self-interest in this case. Here, the intuition is, that since the other players appropriate more than in the SNE, the inequity averse player takes revenge by imposing negative externalities on the other players. The desire to take revenge results from the fact that the large appropriation levels of the others causes disadvantageous inequality for the inequity averse subject. Since appropriating in this area reduces the payoff of the others more than the own payoff, an inequity averse player can reduce the payoff differences. The selfish player on the other hand does not care about payoff differences and therefore appropriates less in this situation.

Third, a part of the inequity averse player's best response lies right *on* the diagonal. This is the area in which symmetric equilibria may exist. There may be equilibria in which subjects appropriate less than in the SNE as well as equilibria in which they appropriate more. Of particular interest are the equilibria to the left of the SNE since in this direction efficiency is increasing (up to the optimal appropriation level of 36). Whether such equilibria do exist depends on the

distribution of the parameters α and β (see our discussion on Proposition 3.2).

Fourth, notice that in case the others do not appropriate at all the best responses of selfish and inequity averse players coincide. At a first glance, this seems counterintuitive, since in a certain sense appropriating nothing is the most friendly choice of the other group members. However, the coincidence of the two best response functions at that point is quite sensible. The reason is, that if the other group members do not appropriate at all, the appropriation decisions of a player does not affect the other players' payoffs at all. This is so because the other players' share of the total revenue $\frac{x_i}{\sum x_j}$ is zero. So why should an inequity averse player not choose the money maximizing appropriation level of 36 units? Remember that the utility function specified in (1) combines a concern for absolute income and for payoff differences. In case the other players do in fact appropriate nothing, utility is equal to $U_i = \pi_i - \beta_i(\pi_i - \pi_{-i}) = \pi_i(1-\beta_i) + \beta_i\pi_{-i}$, where π_{-i} is the individual payoff of each of the $n-1$ other players who appropriate zero. Since π_{-i} is equal to e it does not depend on the choice of player i, and since $\beta_i < 1$, it is clear that even for a highly inequity averse subject money maximizing behavior and utility maximizing behavior coincide.

Figure 2

Given the best response behavior of inequity averse subjects the existence conditions and the nature of symmetric equilibria are described in the next proposition. Note that in this proposition $\min(\beta_i)$ denotes the smallest β_i among all n players and $\min(\alpha_i)$ denotes the smallest α_i.

Proposition 3.2 (Symmetric Equilibria with Inequity Averse Subjects). *There is a symmetric equilibrium in which each subject chooses* $x_i^* = \hat{x}$ *iff* $\hat{x}$ *is in the interval* $\left[\frac{a(1-\min(\beta_i))}{b(1+n(1-\min(\beta_i)))}, \frac{a(1+\min(\alpha_i))}{b(1+n(1+\min(\alpha_i)))}\right]$.

The intuition of Proposition 3.2 is as follows. If both, the smallest α_i and the smallest β_i are equal to zero, the only equilibrium is the SNE, i.e., $\hat{x} = \frac{a}{b(n+1)}$. This means that the presence of only one egoistic player in the group (with $\alpha_i = \beta_i = 0$), suffices to induce all other players to act in a selfish manner, regardless how inequity averse they are. Put differently, a single egoist rules out any efficiency improvement compared to the SNE even if all other $n-1$ players are highly inequity averse.

Proposition 3.2 entails a very strong result. It states that the subject with the 'weakest preferences' for an equitable outcome dictates the outcome for the whole group. Only if the lowest α_i or the lowest β_i are greater than zero asymmetric equilibria that differ from the SNE exist. Of particular interest are equilibria where the smallest β_i is greater than zero. In this case the lower bound of the interval given in Proposition 3.2 is smaller than x_{SNE}, i.e., there are equilibria 'to the left' of the SNE. In these equilibria subjects appropriate less than in the SNE. Similarly, if the smallest α_i is larger than zero, there exist equilibria in which subjects appropriate more than in the SNE. If *all* players in a group are inequity averse and given the parameters of the Ostrom, Gardner and Walker experiment, the range of possible Nash equilibria is $0 < \hat{x} < 9$ where $\hat{x}_{SNE} = 8$ is

always an equilibrium independent of α_i and β_i.[8]

So far we have concentrated on symmetric equilibria. However, there are also asymmetric equilibria. The following proposition provides the details[9].

Proposition 3.3 (Asymmetric Equilibria with Inequity Averse Subjects). *(i) If there are at least k players with $\frac{\beta_i}{\alpha_i} > \frac{n-k}{k-1}$, then there is an equilibrium with less appropriation than in the SNE. In this equilibrium at least k players choose the same appropriation $\widehat{x} < x_{SNE}$; the other players j choose higher appropriation levels. (ii) If there is no k such that there are at least k players with $\frac{\beta_i}{\alpha_i} \geq \frac{n-k}{k-1}$, then there is no equilibrium with less appropriation than in the SNE.*

Corollary 1: If there are $\frac{n}{2}$ or more selfish players, then there is no equilibrium with less appropriation than in the SNE.

The intuition of Proposition 3.3 is straightforward. To get more efficient equilibria than the SNE requires that a relatively large fraction of subjects have rather high $\frac{\beta_i}{\alpha_i}$ combinations. Notice that since $0 \leq \beta_i < 1$ and $\alpha_i \geq \beta_i$, the expression $\frac{\beta_i}{\alpha_i}$ is between zero and one. This means that for $k \leq \frac{n}{2}$ the only equilibrium is the SNE. Only if k is higher than $n/2$ there are $\frac{\beta_i}{\alpha_i}$ combinations to ensure a more efficient equilibrium. For example, if the group size is 8 it takes at least five non-egoistic players to reach such an equilibrium. In this case the $\frac{\beta_i}{\alpha_i}$ combinations of these five subjects must be at least $\frac{\beta_i}{\alpha_i} \geq \frac{8-5}{5-1} = 3/4$. The more people

[8] Since there is a whole range of equilibria the question of equilibrium selection arises. This issue will be discussed in Section 3.3.1.

[9] We restrict attention to the cases where in equilibrium appropriation is less than in SNE.

are non selfish the weaker is the requirement for $\frac{\beta_i}{\alpha_i}$ ($k = 6 \rightarrow \frac{\beta_i}{\alpha_i} \geq \frac{2}{5}; k = 7 \rightarrow \frac{\beta_i}{\alpha_i} \geq \frac{1}{6}; k = 8 \rightarrow \frac{\beta_i}{\alpha_i} \geq 0$). Thus, to reach a more efficient outcome it takes either many subjects with moderate $\frac{\beta_i}{\alpha_i}$-combinations or it takes fewer subjects (but still more than $n/2$), with very high $\frac{\beta_i}{\alpha_i}$-combinations. Notice that the expression $\frac{\beta_i}{\alpha_i}$ rises in β_i and decreases in α_i. It is therefore more likely to reach more efficient outcomes if subjects have a rather large utility loss from advantageous inequality, and a rather small utility loss from disadvantageous inequality.

To summarize the results of this section: What are the prospects for an outcome that is 'better' than that of the SNE? If we look at the requirements of propositions 2 and 3, some scepticism is in place. The requirements are rather tough. In the symmetric case it takes only one selfish subject to ensure that the only equilibrium is SNE. Of course more efficient equilibria are possible. However, we expect that this is the exception rather than the rule. For example, if we assume that there are about 25 percent purely selfish people (a rather 'optimistic' guess) the chance to have no egoist in a randomly drawn group of 8 subjects is about 10 percent. This means that on average in 1 out of 10 groups we would possibly expect less appropriation than in the SNE.

What about the asymmetric case? On a first sight, requirements seem weaker. There are more efficient outcomes even in the presence of a selfish player. However, it takes again more than half of the players who are (i) non selfish and (ii) who have a rather high utility loss from advantageous inequality compared to their loss in utility that derives from disadvantageous inequality. We have strong

doubts that this type of preference is sufficiently frequent. As we have pointed out above we expect that the aversion with respect to disadvantageous inequality is usually much stronger than that arising from advantageous inequality (see also Loewenstein et al., 1989).

Our conclusion is therefore that, even in the presence of many inequity averse and reciprocal subjects, the prospects for achieving a more efficient equilibrium than the SNE are rather weak in the standard CPR-game. This is consistent with much of the reported data according to which - on average - the SNE describes aggregate behavior quite well. In their repeated CPR game Ostrom, Gardner and Walker (1994) report that on average final period appropriation levels in three different groups were 63, 64 and 78 (in case the endowment was 25 tokens) and 60, 63 and 70 (in case the endowment was 10 tokens).[10] These numbers are very close to the standard prediction of 64.

3.2. A CPR Game with Sanctioning Opportunities

In this section we discuss a variant of the standard CPR-game. Again, we follow the experimental setup of Ostrom, Walker and Gardner (1992).[11] Their sanctioning institution is built on the standard CPR-game discussed above. Subjects now first play the standard game and after each round of play all subjects receive data on all individual appropriation decisions. Each subject can then decide to

[10] We concentrate on the behavior of subjects in the final periods to exclude the possible confound of repeated games effects and to make behavior comparable to our one-shot predictions.

[11] A sanctioning institution was first studied by Yamagishi (1986).

sanction any other group member at a certain cost. Technically, any player i can deduct p_{ij} points from player j's payoff at cost cp_{ij} where c is a positive constant smaller than 1. In the reported experiments, different parameter constellations were used where p_{ij} varied from 10 to 80 cents and cp_{ij} varied from 5 to 20 cents.

Since in this type of experiments sanctioning is costly, the standard game theoretic prediction (assuming selfish preferences) is exactly as stated in Proposition 3.1. The rationale for this prediction is straightforward. Why should a rational and selfish person spend resources to sanction another person in the final stage? Since sanctioning is costly and utility depends only on the own material payoff, sanctioning is equivalent to throwing away money. Rational subjects are able to perform the necessary backward induction and everybody therefore knows that nobody will sanction, no matter how egoistical the appropriation behavior on the first stage. Thus appropriation is totally unaffected by the presence of a sanctioning stage.

Contrary to this prediction Ostrom et al. (1994, p. 176) report the following stylized facts.[12]

- Significantly more sanctioning occurs than according to the standard pre-

[12]Moir (1999) studies the impact of monitoring additional to sanctioning. He points out that pure monitoring does not help to overcome excess appropriation. Institutions with a high level of monitoring but a low level of sanctioning may even lead to more appropriation than institutions without any monitoring. A yet different design is suggested by Casari and Plott (1999) where monitoring and punishment is compacted in a single decision. In their treatment efficiency is also higher compared to a baseline treatment without monitoring/punishment.

diction.

- Sanctioning is inversely related to the cost of sanctioning (c).
- Sanctioning is primarily focused on subjects who appropriate the most from the CPR.
- Sanctioning has a modest efficiency enhancing impact on appropriation behavior (i.e., there is less appropriation than in the SNE).
- There is some sanctioning behavior that can be classified as "error, lagged punishment, or 'blind' revenge" (p. 176).
- Taking into account the cost of sanctioning, overall efficiency is similar to the standard CPR without sanctioning opportunities.

As stated above this evidence is largely at odds with the homo oeconomicus perspective. Assuming interdependent preferences, however, this evidence can be explained. In particular our model predicts that defectors will be punished by those players who have sufficiently strong preferences for equity and reciprocity. This punishment serves as a discipline device for the selfish players. As a consequence selfish players have an incentive to act more cooperatively compared to a situation where there is no sanctioning institution. Thus, for a given population (of selfish and inequity averse subjects) the prospects for more efficient appropriation levels are clearly improved in a CPR environment with sanctioning

possibilities. Precise conditions for the existence of equilibria with appropriation levels below the SNE are given in the following proposition.

Proposition 3.4 (Equilibria with Sanctioning Possibilities). *Suppose there is a number $k \leq n$ such that for all players $i \leq k$, the utility parameters α_i and β_i satisfy $c < \frac{\alpha_i}{(1+\alpha_i)(n-k)+(k-1)(1-\beta_i)} = \frac{\alpha_i}{(n-1)(1+\alpha_i)-(k-1)(\alpha_i+\beta_i)} \equiv \widehat{c}$. We call the players who satisfy this condition 'conditionally cooperative enforcers' (CCEs). Suppose further that all players $i > k$ obey the condition $\alpha_i = \beta_i = 0$, i.e., they are selfish. We define $\beta_{\min} = \min_{i \leq k} \beta_i$ as the smallest β_i among the CCEs. Then there is an equilibrium that can be characterized as follows: (i) All players choose $x \in \left[\frac{a(1-\beta_{\min})}{b(1+n(1-\beta_{\min}))}, x_{SNE}\right]$. (ii) If each player does so, there is no sanctioning in the second stage. (iii) If one player chooses a higher appropriation level, then this player is sanctioned equally by all CCEs. The sanctioning equalizes the payoffs of those who sanction and the player who deviates from x. (iv) If more than one player does not play x, an equilibrium of the sanctioning subgame is played.*

Proposition 3.4 determines the critical condition for an equilibrium in which all players appropriate less than x_{SNE}. It states that the cost of sanctioning c must be lower than a certain threshold cost level $\widehat{c}$ which is defined by the preference parameters of the CCEs. The threshold $\widehat{c}$ increases in α_i, β_i and k.

The intuition for the positive relation between $\widehat{c}$ and α_i is the following. A player with a high α_i, experiences a great disutility from disadvantageous inequity. This player is therefore, willing to punish a selfish player who appropriates more

than x (and therefore earns more than player i) even if the sanctioning costs are high.

Why does the critical cost $\widehat{c}$ increase in β_i? Remember that a person with a high β_i has a strong aversion against advantageous inequality. Therefore, such a player i will experience a strong disutility from the advantageous inequity towards the other CCEs when he himself does *not* sanction a selfish player. Put differently, given that the others spend resources in order to sanction a defector, a person with a high β_i will feel solidarity towards these punishers. Thus, while a high α_i leads to punishment because of the inequity towards the deviating person, a high β_i leads to punishment in order to reduce the inequity towards those who actually punish.

Finally, why is $\widehat{c}$ increasing in k? A higher k means that there are more subjects who are willing to punish the players who deviate from x. Therefore, the desire to be in solidarity with those who punish increases as well, i.e., ceteris paribus there will be more punishment.[13]

Notice that when there is a sanctioning opportunity it is much easier to meet the conditions that sustain a cooperative outcome compared to the standard CPR game. The conditions that α_i and β_i have to meet are tougher when sanctioning is ruled out. Put differently, for a given distribution of inequity averse and selfish players it may be impossible to reach an equilibrium with a cooperative

[13]Notice that according to Proposition 3.4, the reason for $\widehat{c}$ to increase in k is *not* because the costs of punishment can be shared between more punishers.

outcome when sanctioning is impossible while there are equilibria with a cooperative outcome when sanctioning is possible. Thus, the model does explain why appropriation is more efficient with a sanctioning device than without. It also correctly predicts that those subjects will be punished who deviate from the 'agreed' appropriation level x. This is a very important point. Since it is exactly the defectors who get punished, the group can discipline the selfish players. As long as there are enough 'norm enforcers', cooperation will be high and stable because the potential deviators face the credible threat of being punished if they behave selfishly. This pattern of punishment behavior can therefore be understood as a norm enforcement device (see Fehr and Gächter 2000a). We would like to add that the Fehr and Schmidt model as well as the Falk and Fischbacher (1999) model share this feature.[14] Finally, the model does explain why sanctioning activities are inversely related to the cost of sanctioning. This follows immediately from Proposition 3.4. For a given set of preferences, the equilibria with cooperative outcomes will be the more likely the lower the cost of punishment.

[14]Other models (e.g., Bolton and Ockenfels, 2000) predict a very different pattern of punishment. In their model punishment is not individually addressed but directed towards a group average. This could imply, e.g., that those who deviate are *not* punished whereas those who cooperate *are* punished. This is totally at odds with the experimental findings. Moreover, it does not take account of the potential of reciprocal or equity preferences to establish and enforce social norms. For a detailed discussion of this point see Falk, Fehr and Fischbacher (2000).

3.3. The Impact of Communication

In all games discussed so far, subjects interact anonymously and without communication. In reality, however, people often communicate. They discuss problems like 'overfishing', make (non-)binding agreements on how to behave and they express approval or disapproval through (face-to-face-)communication. Unless agreements are binding in a strict sense, however, the standard prediction with respect to behavior remains unchanged. When all people are completely selfish there is no hope that after the promise not to appropriate excessively, a subject will actually stick to his promise. When it comes down to give up money just in order to keep one's promise, standard theory predicts that subjects won't hesitate to pursue their material self-interest. In this sense, the opportunity to communicate is irrelevant for the predicted outcomes just as it was in the case with the sanctioning opportunity.

The experimental evidence reported in Ostrom et al. (1994) casts serious doubts at this prediction.[15] They report that subjects "with one and only one opportunity to communicate, obtained an average percentage of net yield above that which was obtained in baseline experiments ... without communication (55 percent compared to 21 percent...)" (p. 198). Allowing subjects to communicate repeatedly increases efficiency even more (73 percent).[16]

In a meta study by Sally (1995) on the determinants of cooperative behavior

[15] See also Ostrom and Walker (1991).

[16] On communication see also the paper by Kopelman, Weber and Messick in this volume.

in more than 100 public goods experiments, communication has a significant and positive influence. In one-shot games, cooperation is raised by about 45 percent on average, whereas in repeated games the increase is 40 percent. Communication, however, is an elusive term. In some experiments subjects really talk to each other, i.e., they exchange verbal and facial expressions. In other experiments, subjects do not communicate face-to-face but via computer or written notes. In yet other experiments, subjects do not actually talk but simply identify each other, i.e., they do not exchange any verbal information. As diverse as the experiments is the discussion on why communication has a positive impact. Kerr and Kaufman-Gilliland (1994), e.g., discuss nine different effects communication may have. It is not our aim to address these issues at length. The purpose of this section is to discuss how communication affects decisions when fairness concerns play a role. The two main effects we consider as relevant from this perspective are coordination and the expression of approval and disapproval. Both effects were possibly in the CPR communication treatment outlined above since subjects could not only exchange information but also did see and talk to each other.

3.3.1. Coordination

Remember that in the standard CPR game with inequity averse or reciprocal preferences there may be multiple equilibria. The SNE is always one of these equilibria, among others which are more (or less) efficient. To emphasize our point let us abstract from all details and assume a two player CPR game. In

terms of material payoffs the game could look like the one expressed in Figure 3a, i.e., the CPR game is similar to a Prisoner's Dilemma. Even though it is in their common interest to choose the low appropriation level, both players can individually improve their material payoffs if they choose the high appropriation strategy. This yields the unique equilibrium where both players choose the high appropriation. If both players have purely selfish preferences, this is the end of the story (and communication has no impact).

In the presence of reciprocal preferences, however, the CPR game is no longer a Prisoner's Dilemma (see Figure 3b). The reason is that if both players are sufficiently reciprocally motivated they don't like to cheat on the other player. If, e.g., player 1 chooses the low appropriation strategy, a reciprocal player 2 is better off choosing the low instead of the high appropriation level and vice versa. Even though players forgo some material payoffs they have a higher utility if they reciprocate the nice behavior of the other player. If player 1 chooses the high appropriation level, however, player 2 has no desire to choose the low level (neither if he is a selfish nor if he is a reciprocal player). Instead, player 2 will in this case also play the high appropriation strategy. As a consequence, there are two (pure) equilibria now, the efficient equilibrium with low appropriations and the inefficient one with high appropriations. Put differently, the Prisoner's Dilemma game in Figure 3a with a unique and inefficient equilibrium has turned into a coordination game with one efficient and one inefficient equilibrium. Game theory does not help much in this situation. It simply predicts that *some* Nash

equilibrium will be played, but not which one[17].

In the presence of multiple equilibria subjects face a tremendous strategic uncertainty. How shall a person know which strategy the other player will select? It is obvious that communication can have a positive impact in a situation of strategic uncertainty. In fact it has been shown experimentally that communication can help players to coordinate on better equilibria[18]. As an example take Cooper et al. (1992) who study different coordination games with and without communication. They find that depending on the precise structure of the coordination game either one-way or two-way communication clearly improves efficiency. This holds even though all announcements are non-binding.

In the CPR experiments with communication mentioned above, players had intensive opportunities for communication since they could actually talk to each other and were (with some restrictions) allowed to discuss anything they wanted. As reported in Ostrom et al. (1994) subjects usually came to the agreement to appropriate a particular amount (e.g., 5 tokens). If this is the case, coordination on 'good' equilibria seems possible. Given these extensive communication opportunities and given the fact that there is usually a substantial fraction of reciprocal subjects, it seems therefore quite likely that communication raised efficiency because subjects could coordinate their choices on more efficient equilibria.

[17] See, however, the literature on equilibrium selection, e.g., Harsanyi and Selten (1988).

[18] On coordination game experiments see Ochs (1995).

3.3.2. Communication as a Sanctioning Device

Social interactions are frequently associated with social approval or disapproval. The anticipation of such social rewards and punishments may have important economic consequences. For example, it may affect the efficiency of team production and the decisions in diverse areas such tax evasion, the exploitation of the welfare state, criminal activities, union membership, and voting behavior. The behavioral role of social rewards and punishments is stressed in social exchange theory (Blau, 1964). In contrast to pure economic exchanges social exchanges involve not only the exchange of economic rewards but also the exchange of social rewards. The admiration or the contempt that is sometimes expressed by parents, teachers, professional colleagues and spectators are prime examples of a social reward. In general social rewards are not based on explicit contractual arrangements but are triggered by spontaneous positive or negative emotions which can be interpreted as approval and disapproval, respectively.

Of course approval as well as disapproval can be communicated and can have an important impact on individual behavior in a CPR game. People who talk to each other enter a social relationship. Within this relationship, exchange of approval and disapproval is possible. Two assumptions must be met in order to observe more cooperative behavior compared to SNE, however. *First*, there must be subjects who actually care about approval or disapproval and who change their behavior in the expectation of such approval or disapproval. *Second*, there must be subjects who actually express approval or disapproval. The first condition is

obvious. The second condition is important since it is usually not costless to express approval and in particular disapproval. Our point is that *reciprocally motivated subjects* are willing to bear the cost and to reciprocate the cooperative or non-cooperative actions by others. Thus, preferences as assumed in our model may explain why communication in combination with the expression of approval and disapproval can have a positive impact on cooperative behavior.

Taken together we have described two potential channels through which communication may effect cooperative behavior in the presence of reciprocal preferences. While the first rests only on the exchange of information the second is built on the possibility of communication face to face. We would therefore expect that communication effects are particularly strong if face to face communication is possible (as it is the case in the treatment discussed above). This is also the conclusion of Rocco and Wargelein 1995 who report a study showing that it is the communication face to face which makes the big difference (on this point see also Ostrom (1998)).

4. Public Goods: A Comparison

So far we have analyzed CPR games. However, many of the arguments apply also to public goods games. In fact, public goods games and CPR games are very similar. Whereas in a CPR game, subjects' decisions impose negative externalities on other subjects, subjects in a public goods game produce positive externalities. In a CPR game it is nice or kind not to appropriate too much while in a public

goods game it is kind not to contribute too little to the public good. Public good situations are very important and very frequent in reality. Moreover, there exists a huge experimental literature on public goods games. As we will show, many of the findings reported on CPR problems carry over to those of public goods. In the following we discuss a one-stage public goods game (similar to the standard CPR game) and a two-stage public goods game, where after the first stage, subjects have a sanctioning opportunity (similar to the CPR with sanctioning opportunities).

We start with the following linear public good game. There are $n \geq 2$ player who decide simultaneously on their contribution levels $g_i \in [0, y]$, $i \in [1, ..., n]$, to the public good. Each player has an endowment of y. The monetary payoff of player i is given by $x_i(g_1, ..., g_n) = y - g_i + a \sum_j g_j$ where $1/n < a < 1$. Since $a < 1$, a marginal investment into G causes a monetary loss of $(1 - a)$, i.e., the dominant strategy of a completely selfish player is to choose $g_i = 0$. However, since $a > 1/n$, the aggregate monetary payoff is maximized if each player chooses $g_i = y$.

Consider now a slightly different public good game that consists of two stages. At stage 1 the game is identical to the previous game. At stage 2 each player i is informed about the contributions of all other players and can simultaneously impose a *costly punishment* on the other players, just as in the sanctioning CPR game discussed above.

What does the standard model predict for the two-stage game? Since punishments are costly, players' dominant strategy at stage two is to not punish.

Therefore, if selfishness and rationality are common knowledge, each player knows that the second stage is completely irrelevant. As a consequence, players have exactly the same incentives at stage 1 as they have in the one-stage game without punishments, i.e., each player's optimal strategy is to contribute nothing.

To what extent are these predictions of the standard model consistent with the data from public good experiments? For the one-stage-game there are, fortunately, a large number of experimental studies. In a meta study of 12 experimental studies (with a total of 1042 subjects participating) Fehr and Schmidt (1999) report that in the final period of public goods games without punishment the vast majority of subjects plays the equilibrium strategy of complete free-riding. On average 73 percent of all subjects choose $g_i = 0$ in the final period. It is also worth mentioning that in addition to those subjects who play exactly the equilibrium strategy there is very often a non negligible fraction of subjects who play "close" to the equilibrium. In view of the facts it seems fair to say that the standard model "approximates" the choices of a big majority of subjects rather well. However, if we turn to the public good game with punishment there emerges a radically different picture although the standard model predicts the same outcome as in the one-stage game. Figure 4 shows the distribution of contributions in the final period of the two-stage game conducted by Fehr and Gächter (2000a). Note that the same subjects generated the distribution in the game without and in the game with punishment. Whereas in the game without punishment most subjects play close to complete defection a strikingly large fraction of 82.5 percent

cooperates fully in the game with punishment. Fehr and Gächter report that the vast majority of punishments is imposed by cooperators on the defectors and that lower contribution levels are associated with higher received punishments. Thus, defectors do not gain from free-riding because they are being punished.

Figure 4

When these results are compared with the evidence from CPR games a striking similarity arises. In the standard CPR game average behavior (in final periods) is pretty consistent with the standard prediction. However, if subjects have the opportunity to sanction each other, behavior becomes much more cooperative - even though the standard prediction yields the same outcome. As we have seen in our discussion above, our fairness model can explain the evidence in both CPR games. This holds also for the public goods games. The intuition for the one-stage public good game is straightforward. Only if sufficiently many players have a dislike for an advantageous inequity they can possibly reach some cooperative outcome. As long as there are only a few players who are willing to contribute if others contribute as well, they would suffer too much from the disadvantageous inequality caused by the free-riders. Thus, inequity averse players prefer to defect if they know that there are selfish players. To put it differently: The greater the aversion against being the sucker the more difficult it is to sustain cooperation in the one-stage game.

Consider now the public good game with punishment. To what extent is our model capable of accounting for the very high cooperation in this treatment? The

crucial point is that free-riding generates a material payoff advantage relative to those who cooperate. Since $c < 1$, cooperators can reduce this payoff disadvantage by punishing the free-riders. Therefore, if those who cooperate are sufficiently upset by the inequality to their disadvantage, i.e., if they have sufficiently high α's, then they are willing to punish the defectors even though this is costly to themselves. Thus, the threat to punish free-riders may be credible, which may induce potential defectors to contribute at the first stage of the game.

Notice that according to the present model (and the inequity aversion approach in general), a person will punish another person if and only if this reduces the inequity between the person and his opponent(s). Therefore, as long as $c < 1$ (as it is the case in the CPR problem and the public goods game analyzed above) the model predicts punishments for sufficiently inequity averse subjects. If, on the other hand, $c \geq 1$ the Fehr-Schmidt model predicts no punishment at all. This holds regardless whether we look at public goods games or at the CPR problems. Experimental evidence suggests, however, that many subjects in fact punish others even if punishment does *not* reduce inequity (as it is the case if $c \geq 1$). Falk, Fehr and Fischbacher (2000) present several experiments which address this question in more detail. As it turns out, a substantial amount of punishment occurs even in situations where inequity cannot be reduced. For example, in one of their public goods games with punishment they implemented a punishment cost of $c = 1$. Nevertheless, 46.8 percent of the subjects who cooperated in this game punished defectors. The conclusion from Falk, Fehr and Fischbacher (2000) is, therefore,

that the desire to reduce inequity cannot be the only motivation to punish unkind acts. An alternative interpretation is offered in Falk and Fischbacher (1999) who model punishment as the desire to reduce the unkind players' payoff(s). Their model also correctly predicts punishments in those situations where punishment is costly and cannot reduce inequity.

5. Discussion

In the preceding sections we have demonstrated that with the help of a simple fairness theory many stylized facts of CPR or public good experiments can be explained. In fact, the range of experiments which have successfully analyzed with the help of our fairness theories is even much wider. Both, the model by Fehr and Schmidt (1999) as well as that of Falk and Fischbacher (1999) are capable of predicting correctly a wide variety of seemingly contradictory experimental facts. They are, in particular, capable of reconciling the puzzling evidence that in competitive experimental markets with complete contracts very unfair outcomes, that are compatible with the predictions of the pure self-interest model, can emerge, while in bilateral bargaining situations or in markets with incomplete contracts stable deviations from the predictions of the self-interest model, in the direction of more fair and equitable outcomes, are the rule.

The basic behavioral principle which is formalized in the models is that a substantial fraction of the subjects acts conditionally on what other subjects do. If others are nice or cooperative they act cooperatively as well while if others are

hostile they retaliate. The models also pay attention to the fact that there are large individual differences between subjects. In particular it is assumed that there are selfish subjects who behave in the way predicted by standard economic theory and reciprocal subjects who exhibit the type of conditional behavior just mentioned. *The interaction of these diverse motivations and the institutional set-up is responsible for the observed experimental outcomes.* If there is no institutional rule which externally enforces cooperation or that allows for sanctioning possibilities, the interaction of selfish and conditional subjects frequently leads to non-cooperative outcomes. If on the other hand subjects dispose of sanctioning possibilities, the reciprocal subjects are able to discipline selfish players. As a consequence more cooperative outcomes will emerge. This approach goes beyond the standard economic conception not least because it assigns institutions a much more important role. In the presence of reciprocal and selfish subjects institutions determine which type of preference is pivotal for the equilibrium outcome. In a sense institutions select the type of player that shapes the final result.

There are of course several important behavioral factors which we have not addressed or which cannot be explained with the help of the presented theoretical framework. The Kopelman, Weber and Messick paper (in this volume) summarizes a broad variety of influential factors which are beyond the present scope of our theory. While our models speak to factors like social motives, trust, payoff structure and group size they are mute with respect to gender, culture frames, etc. The models presented in this paper also do not ask about possible evolutionary

roots of reciprocal preferences or pro-social behavior in general. This important question is addressed in Richerson, Boyd and Paciotti (this volume).[19]

Another remark is also in place. We have emphasized the importance of reciprocity and inequity aversion but have not mentioned the impact of reputation and repeated game effects. Many of the real life CPR- or public goods problems are in fact 'played' repeatedly. In such repeated interaction players usually can condition their behavior on past behavior of others. This allows players to build up reputations and to ensure cooperative outcomes even among selfish players. In the parlance of game theory this kind of cooperation may be supported as an equilibrium in infinitely repeated games (folk theorems) or in finitely repeated games with incomplete information (see Kreps, Milgrom, Roberts, and Wilson (1982)[20]). Many experiments have demonstrated the efficiency enhancing effect of repeated vs. one-shot interactions. Moreover, it has been shown that reciprocity and repeated game effects interact in a complementary way (Gächter and Falk 2000). In their experimental study of a bilateral labor relation the reciprocal relationship between workers and firms is significantly increased in a repeated interaction compared to one-shot encounters. The driving force behind this "crowding in" of

[19] See also Sethi and Somananthan (2000) and Huck and Oechssler (1999). See also de Waal (1996) who shows that conditional behavior is observed among chimpanzees. Their food sharing behavior exhibits some reciprocal pattern: A chimpanzee is ceteris paribus more willing to share food with another chimpanzee if the latter has shared with the former in the past.

[20] Notice that the latter model is built on the assumption that there exist selfish *and* reciprocal (tit-for-tat) types.

reciprocal behavior is the fact that people who behave selfishly in the one-shot game have an incentive to imitate reciprocity in the repeated game.[21] Thus, in the presence of repeated game incentives, the prospects for cooperative outcomes are expected to be better than according to the one-shot analysis undertaken in this paper.

[21] On the complementary relationship between reputation and reciprocity see also the paper by Ostrom (1998).

References

Adams, J.S.

1963 Toward an Understanding of Inequity. *Journal of Abnormal and Social Psychology* 62:422-436.

Agell, J., and P. Lundborg

1995 Theories of Pay and Unemployment: Survey Evidence from Swedish Manufacturing Firms. *Scandinavian Journal of Economics* 97:295 - 307.

Berg, J., J. Dickhaut, and K. McCabe

1995 Trust, Reciprocity and Social History. *Games and Economic Behavior* 10:122-142.

Bewley, T.

1998 Why not cut Pay? *European Economic Review* 42:459-490.

Blau, P.

1964 *Exchange and Power in Social Life*. New York: Wiley.

Blount, S.

1995 When Social Outcomes aren't Fair: The Effect of Causal Attributions on Preferences. *Organizational Behavior and Human Decision Processes* 63(2):131-144.

Bolton, G.E., and A. Ockenfels

2000 A Theory of Equity, Reciprocity and Competition. *American Economic Review* 90:166-193.

Richerson, P., R. Boyd, and B. Paciotti,

2000 An Evolutionary Theory of Commons Management. *This volume*.

Casari, M., and C. Plott

1999 Agents Monitoring Each Other in a Common-Pool Resource Environment. Working paper, California Institute of Technology, Pasadena.

Charness, G., and M. Rabin

2000 *Social Preferences: Some Simple Tests and a new Model*. Working paper, University of Berkeley.

Clark, A.E., and A.J. Oswald

1996 Satisfaction and Comparison Income. *Journal of Public Economics* 61:359-381.

Cooper, R., D. DeJong, R. Forsythe and T. Ross

1992 Communication in Coordination Games. *Quarterly Journal of Economics* 107:739-771.

Davis, J.A.

1959 A Formal Interpretation of the Theory of Relative Deprivation. *Sociometry* 102:280-296.

Davis, D., and Ch. Holt

1993 *Experimental Economics*. Princeton: Princeton University Press.

Dawes, R.M., and R. Thaler

1988 Cooperation. *Journal of Economic Perspectives* 2(3):187-197.

Dufwenberg, M., and G. Kirchsteiger

1998 A Theory of Sequential Reciprocity. Discussion paper, CentER, Tilburg University.

Falk, A., E. Fehr, and U. Fischbacher

2000 Informal Sanctions. Working paper, Institute for Empirical Research, University of Zurich.

Falk, A., and U. Fischbacher

1998 A Theory of Reciprocity. Working paper 6, Institute for Empirical Research, University of Zurich.

Falk, A., and M. Knell

2000 Choosing the Joneses: On the Endogeneity of Reference Groups? Working paper 53, Institute for Empirical Research, University of Zurich.

Fehr, E., G. Kirchsteiger, and A. Riedl

1993 Does Fairness prevent Market Clearing? An Experimental Investigation. *Quarterly Journal of Economics* 108:437-460.

Fehr, E., and S. Gächter

2000a Cooperation and Punishment in Public Good Experiments - An Experimental Analysis of Norm Formation and Norm Enforcement. *American Economic Review* forthcoming.

2000b Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives* forthcoming.

Fehr, E., and K. Schmidt

1999 A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 114:817-851.

Festinger, L.

1954 A Theory of Social Comparison Processes. *Human Relations* 7:117-140.

Frank, R.H.

1985 *Choosing the Right Pond - Human Behavior and the Quest for Status*. Oxford: Oxford University Press.

Gächter, S., and A. Falk

1999 Reputation or Reciprocity?, Working paper 19, Institute for Empirical Research, University of Zurich.

Gordon, S.

1954 The Economic Theory of Common-Property Resource: The Fishery. *Journal of Political Economy* 62:124-142.

Güth, W., R. Schmittberger, and B. Schwarze

1982 An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior and Organization* 3(3):367-88.

Hardin, G.

1968 The tragedy of the Commons. *Science* 162:1243-1248.

Harsanyi, J., and R. Selten

1988 *A General Theory of Equilibrium Selection in Games*. Cambridge: MIT Press.

Homans, G.C.

1961 *Social Behavior: Its Elementary Forms*. New York: Harcourt, Brace & World.

Huck, S., and J. Oechssler

1996 The Indirect Evolutionary Approach to Explaining Fair Allocations. *Games and Economic Behavior* 28:13-24.

Kerr, N., and C. Kaufmann-Gilliland

1994 Communication, Commitment and Coordination in Social Dilemmas. *Journal of Personality and Social Psychology* 66:513-529.

Kopelman, S., J. M. Weber, and D. M. Messick

2000 Commons Dilemma Management: Recent Experimental Results. *This Volume*.

Levine, D.

1998 Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics* 1:593-622.

Loewenstein, G.F., L. Thompson, and M.H. Bazerman

1989 Social Utility and Decision Making in Interpersonal Contexts. *Journal of Personality and Social Psychology* 57:426-441.

Moir, R.

1999 Spies and Swords: Behavior in Environments with Costly Monitoring and Sanctioning. Working paper, University of New Brunswick.

Ochs, J.

1995 Coordination Problems. In *Handbook of Experimental Economics*, J. Kagel and A. Roth, eds. Princeton: Princeton University Press.

Ostrom E.

1990 *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

1998 A Behavioral Approach to the Rational Choice Theory of Collective Action - Presidential Address of the American Political Science Association 1997. *American Political Science Review* 92:1-22.

Ostrom, E., R. Gardner, and J. Walker

1994 *Rules, Games, and Common Pool Resources*. Michigan: The University of Michigan Press.

Ostrom, E., and J. Walker

1991 Communication in a Commons: Cooperation without External Enforcement. In *Laboratory Research in Political Economy*, T. Palfrey, ed. Ann Arbor: University of Michigan.

Ostrom, E., J. Walker, and R. Gardner

1992 Covenants with and without a Sword: Self-Governance is Possible. *American Political Science Review* 40:309-317.

Pollis, N.P.

1968 Reference Groups Re-examined. *British Journal of Sociology* 19:300-307.

Rabin, M.

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

Rocco, E., and M. Warglien

1995 Computer Mediated Communication and the Emergence of Electronic Opportunism. Working paper RCC 13659, Universita degli Studi di Venezia.

Runciman, W.G.

1966 *Relative Deprivation and Social Justice*. New York: Penguin.

Sethi R., and E. Somananthan

2000 Preference Evolution and Reciprocity. *Journal of Economic Theory* forthcoming.

de Waal, F.

1996 *Good Natured. The Origins of Right and Wrong in Humans and Other Animals*. Harvard: Harvard University Press.

Walker, J., R. Gardner, and E. Ostrom

1990 Rent Dissipation in a Limited Access Common-Pool Resource: Experimental Evidence. *Journal of Environmental Economics and Management* 19:203-211.

Yamagishi, T.

1986 The Provision of a Sanctioning System as a Public Good. *Journal of Personality and Social Psychology* 51:110-116.

Appendix

Proof Proposition 3.1 (Selfish Nash Equilibrium)

The standard CPR game we look at has the following form.

$$\pi_i = 50 - 5x_i + \left[\frac{x_i}{\sum x_j}\right][23\sum x_j - .25(\sum x_j)^2]$$

To ease notation, we transform the latter equation to get

$$\pi_i = 50 + 18x_i - 0.25x_i \sum x_j$$
$$\pi_i =: e + ax_i - bx_i \sum x_j.$$

To find the selfish best reply for x_i,we calculate

$$\frac{\partial \pi_i}{\partial x_i} = a - 2bx_i - b\sum_{j \neq i} x_j$$

Setting it equal to zero, we get as best reply $x_i = \max(0, \frac{1}{2}(\frac{a}{b} - \sum_{j \neq i} x_j))$

First suppose that all $x_i^* > 0$. In this case

$$2bx_i^* = a - b\sum x_j^*$$

Summing up the terms for all i, we get $2b\sum x_i^* = na - bn\sum x_j^*$. Hence $\sum x_i^* = \frac{na}{b(n+1)}$. Entering the sum into $2bx_i^* = a - b\sum x_j^*$, we get $x_i^* = \frac{a}{b(n+1)}$. Now consider that there are some players who choose 0. Let n_0 be the number of players who choose x_i equal to zero. Then all values above zero must be equal to $\frac{a}{b(n-n_0+1)}$. Calculating the best reply for one of the n_0 players who originally played 0 now yields a contradiction. QED

Social optimum: To find the social optimum, we calculate

$$\frac{\partial\left(\sum \pi_i\right)}{\partial x_i} = \frac{\partial\left(e + a x_i - b x_i \sum x_j\right)}{\partial x_i} = a - 2b \sum_j x_j$$

Hence, in the social optimum, we get $\sum_j x_j = \frac{a}{2b}$.

Proof of Proposition 3.2 (Symmetric Equilibria with Inequity Averse Subjects)

First note that if $\sum x_j < \frac{a}{b}$, the players who choose the higher appropriation level have also higher payoffs:

$$\pi_j - \pi_i = (e + a x_j - b x_j \sum x_k) - (e + a x_j - b x_j \sum x_k) = (a - b \sum x_k)(x_j - x_i)$$

So let us first consider this case.

Suppose, all players $j \neq i$ choose $x_j = \widehat{x} \leq x_{SNE}$. Since U_i is concave in x_i, the best reply is unique. So, to show that $x_i^* = \widehat{x}$ is the best reply, it is sufficient to show that it is a *local* optimum. It is clear that $x_i^* \geq \widehat{x}$ since otherwise player i could improve his material payoff as well as he could reduce inequity by increasing x_i. It remains to check that there is no incentive for i to increase x_i above $\widehat{x}$. As the following calculation shows, the derivative from above $\frac{\partial U_i}{\partial x_i^+}(x)$ is a linear function in x. So, player i has no incentive to increase x_i^* above $\widehat{x}$ iff this derivative equals at least zero.

$$\begin{aligned}
0 \quad \geq \quad & \frac{\partial U_i}{\partial x_i^+} = \frac{\partial}{\partial x_i}\left(\pi_i - \frac{\beta_i}{n-1} \sum_{j, \pi_i > \pi_j} (\pi_i - \pi_j)\right) = \\
= \quad & \frac{\partial}{\partial x_i}\left(\pi_i - \beta_i\left(\pi_i - \pi_j\right)\right) = \\
= \quad & \frac{\partial}{\partial x_i}\left(\pi_i(1 - \beta_i) + \beta_i \pi_j\right) =
\end{aligned}$$

$$= \quad (a - bx_i - b\sum x_j)(1 - \beta_i) - \beta_i b x_i$$

Thus, we get a critical condition for x_i^*

$$x_i^* \geq \frac{a(1-\beta_i)}{b(1+n(1-\beta_i))} \tag{6.1}$$

The right hand side of this inequality is decreasing in β_i. Thus, the left inequality of the proposition is satisfied, if and only if 6.1 is satisfied for all i.

Assume now all other players $j \neq i$ choose $x_j = \widehat{x} > x_{SNE}$; $\widehat{x} < \frac{a}{nb}$. Now, the critical condition is $\frac{\partial U_i}{\partial x_i^-}(x) > 0$.

$$\begin{aligned}
\frac{\partial U_i}{\partial x_i^-} &= \frac{\partial}{\partial x_i}\left(\pi_i - \frac{\alpha_i}{n-1}\sum_{j,\pi_j>\pi_i}(\pi_j - \pi_i)\right) = \\
&= \frac{\partial}{\partial x_i}\left(\pi_i - \alpha_i\left(\pi_j - \pi_i\right)\right) = \\
&= \frac{\partial}{\partial x_i}\left(\pi_i(1+\alpha_i) - \alpha_i\pi_j\right) = \\
&= (a - bx_i - b\sum x_j)(1+\alpha_i) + \alpha_i b x_i
\end{aligned}$$

Thus, we get a critical condition for x_i^*

$$x_i^* \leq \frac{a(1+\alpha_i)}{b(1+n(1+\alpha_i))} \tag{6.2}$$

The right hand side of this inequality is increasing in α_i. Thus, the left inequality of the proposition is satisfied, if and only if 6.2 is satisfied for all i and we get the right inequality in the proposition.

It remains to show that there is no equilibrium with appropriation decisions above $\frac{a}{nb}$. We fix $x_j = \widehat{x} > \frac{a}{nb}$. The critical condition is $\frac{\partial U_i}{\partial x_i^-}(x) \geq 0$. Since a decrease of the appropriation level now generates inequity in favor of player i, we get the following condition.

$$\begin{aligned}
0 \quad \leq \quad & \frac{\partial U_i}{\partial x_i^-} = \frac{\partial}{\partial x_i}\left(\pi_i - \frac{\beta_i}{n-1}\sum_{j,\pi_i>\pi_j}(\pi_i - \pi_j)\right) = \\
= \quad & (a - 2bx_i - b(n-1)x^*)(1-\beta_i) - \beta_i b x_i \leq \\
\leq \quad & (a - 2bx_i - b(n-1)\frac{a}{nb})(1-\beta_i) - \beta_i b x_i = \\
= \quad & (\frac{a}{nb} - 2x_i)b(1-\beta_i) - \beta_i b x_i \leq \\
\leq \quad & (\widehat{x} - 2x_i)b(1-\beta_i) - \beta_i b x_i
\end{aligned}$$

Because $\beta_i < 1$, the last term is negative if x_i is close to $\widehat{x}$. Hence, there are no equilibria with $x^* > \frac{a}{nb}$. QED

Proof of Proposition 3.3 (Asymmetric Equilibria with Inequity Averse Subjects)

We first show (ii): Let us assume that there is an equilibrium with some $x_i^* < x_{SNE}$. By reordering the players, we can assume that we have $x_1^* \leq x_2^* \leq ... \leq x_n^*$. Furthermore let k be the highest index for which $x_1^* = x_k^*$. Now let's consider $i \leq k$. Because we are in an equilibrium, we have $\frac{\partial U_i}{\partial x_i^+} \leq 0$. Remember that $\pi_j - \pi_i = (a - b\sum x_k)(x_j - x_i)$. So

$$0 \quad \geq \quad \frac{\partial U_i}{\partial x_i^+}(x^*) = \frac{\partial}{\partial x_i}\left(\pi_i - \frac{(a - b\sum x_k)}{N-1}\left(\beta_i(k-1)(x_i - x_1^*) + \alpha_i\sum_{j>k}\left(x_j^* - x_i\right)\right)\right) =$$

$$
\begin{aligned}
= \quad & \frac{\partial \pi_i}{\partial x_i} - \frac{(a - b\sum x_k)}{N-1}\left(\beta_i(k-1) - \alpha_i(N-k)\right) + \\
& + \frac{b}{N-1}\left(\beta_i(k-1)(x_i - x_1^*) + \alpha_i \sum_{j>k}\left(x_j^* - x_i\right)\right) \geq \\
\geq \quad & -\frac{(a - b\sum x_k)}{N-1}\left(\beta_i(k-1) - \alpha_i(N-k)\right)
\end{aligned}
$$

Hence

$$\beta_i(k-1) - \alpha_i(N-k) \geq 0$$

or

$$\frac{\beta_i}{\alpha_i} \geq \frac{n-k}{k-1}$$

which proves (ii).

Let us now come to the proof of (i). Assume without loss of generality that for i between 1 and k, we have $\frac{\beta_i}{\alpha_i} > \frac{n-k}{k-1}$. This implies $k > \frac{n}{2}$ because $\frac{\beta_i}{\alpha_i} < 1$. We will show that there is an equilibrium with $x_1 = x_2 = ... = x_k < x_{SNE}$. For $x \in [0, x_{SNE}]$ we define the strategy combination $s(x)$ as follows: We fix $s(x)_i = x$ for $i \leq k$ and choose $s(x)_j$ for $j > k$ as the joint best reply. That means that $s(x)_j$ is a part of a Nash equilibrium in the $(n-k)$-player game induced by the fixes choice of x by the first k players. Because at least half of the players choose x, the best reply can never be smaller than x (by increasing the appropriation level below x, the material payoff could be increased and the inequity disutility could be decreased as well). If we find $\hat{x}$, such that $\frac{\partial U_i}{\partial x_i^+}(s(\hat{x})) \leq 0$ for $i \leq k$, then $(\hat{x}, ..., \hat{x}, s(\hat{x})_{k+1}, s(\hat{x})_n)$ is the desired equilibrium.

Now

$$\begin{aligned}\frac{\partial U_i}{\partial x_i^+}(s(\hat{x})) &= \frac{\partial \pi_i}{\partial x_i} - \frac{(a - b\sum s(\hat{x})_j)}{N-1}\left(\beta_i(k-1) - \alpha_i(N-k)\right) + \\ &\quad + \frac{b}{N-1}\left(\beta_i(k-1)(x_i - \hat{x}) + \alpha_i \sum_{j>k}(s(\hat{x})_j - x_i)\right) \\ &= \frac{\partial \pi_i}{\partial x_i} - \frac{(a - b\sum s(\hat{x})_j)}{N-1}\left(\beta_i(k-1) - \alpha_i(N-k)\right) + \\ &\quad + \frac{b}{N-1}\left(\alpha_i \sum_{j>k}(s(\hat{x})_j - x_i)\right)\end{aligned}$$

We then get

$$\lim_{\hat{x}\to x_{NE}} \frac{\partial U_i}{\partial x_i^+}(s(\hat{x})) = \frac{(a - b\sum s(\hat{x})_j)}{N-1}\left(\beta_i(k-1) - \alpha_i(N-k)\right) < 0$$

Hence, for some $\hat{x}$ near enough to x_{SNE}, we get $\frac{\partial U_i}{\partial x_i^+}(s(\hat{x})) < 0$. The strategy combination $s(\hat{x})$ is the desired equilibrium. QED

Proof of Proposition 3.4 (Equilibria with Sanctioning Possibilities)

Proof: We note

(A) The condition $x \in \left(\frac{a(1-\beta_{\min}))}{b(1+n(1-\beta_{\min})))}, x_{SNE}\right]$ guarantees that x maximizes the utility for the CCEs if all other players choose x.

We call a player *deviator* who chooses an appropriation x' that results in a higher payoff in the first stage compared to choosing x.

(B) If there is a single deviator, then the payoffs for the other players are smaller compared to the situation where there is no deviator.

First, if punishment is executed, the selfish players have no incentive to deviate. Since punishment results in equal payoffs for the CCEs and for the deviator, this

payoff is smaller than the payoff in the first stage of the CCE. Hence, a selfish deviator has no incentive to deviate if he risks to be punished.

Let us now prove that no CCE has an incentive to change the punishment strategy if a selfish player has not chosen x. Let π_P be the payoff after punishment for the CCEs and for the deviator. Let π_S be the payoff of the selfish players. A CCE player has never an incentive to choose a higher punishment than the equilibrium punishment. This only increases inequity with respect to all players and reduces the material payoff. So let w be a positive number and assume CCE player i chooses a punishment of $p - w$ instead of p. We get

$$\begin{aligned} U_i &= \pi_P + cw - \frac{(n-k-1)\alpha_i}{n-1}\left(\pi_S - (\pi_P + cw)\right) - \frac{\alpha_i}{n-1}\left(\pi_p + w - (\pi_p + cw)\right) - \\ &\quad - \frac{(k-1)\beta_i}{n-1}\left(\pi_p + cw - \pi_p\right) \end{aligned}$$

This is a linear function in w. Player i has no incentive to deviate iff the derivative with respect to w is negative, so iff

$$\begin{aligned} 0 \quad &\geq \quad \frac{\partial U_i}{\partial w} = c + c\frac{(n-k-1)\alpha_i}{n-1} - (1-c)\frac{\alpha_i}{n-1} - c\frac{(k-1)\beta_i}{n-1} \\ &\iff \quad c[(n-1) + (n-k-1)\alpha_i + \alpha_i - (k-1)\beta_i] \leq \alpha_i \\ &\iff \quad c \leq \frac{\alpha_i}{(n-1)(1+\alpha_i) - (k-1)(\alpha_i + \beta_i)} \end{aligned}$$

QED.

Figure 1: Preferences of inequity aversion

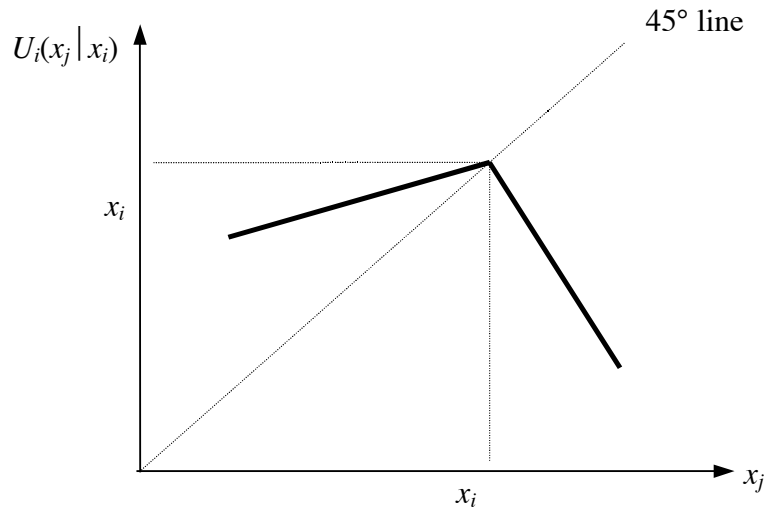

Figure 2: Best response behavior in a standard CPR-game (alpha = 4, beta = 0.6)

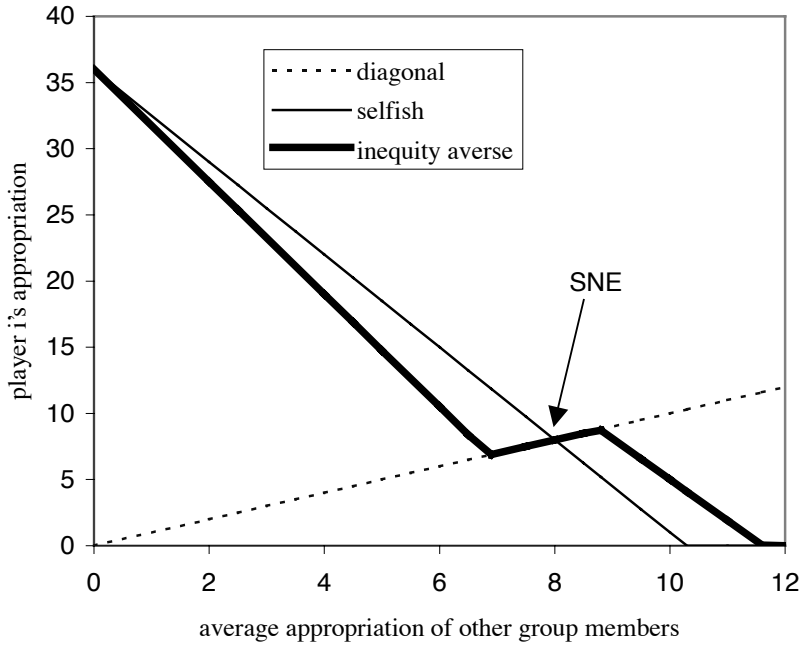

Figure 3a: A simple CPR-game without reciprocal preferences

		Player 2	
		low appropriation	high appropriation
Player 1	low appropriation	10,10	0,15
	high appropriation	15,0	5,5

Figure 3b: A simple CPR-game in the presence of reciprocal preferences

		Player 2	
		low appropriation	high appropriation
Player 1	low appropriation	10,10	0,9
	high appropriation	9,0	5,5

Figure 4: Contributions to the public good with and without punishment