The Design of (De)centralized Punishment Institutions for Sustaining Cooperation*

Michael Kosfeld† Arno Riedl‡

January 27, 2004

1 Introduction

Looking at human society, it seems fair to say that one of the most fundamental conflicts — both from an economic and social viewpoint — arises when social welfare maximization and individual utility maximization are at odds. In the presence of these conflicting interests the otherwise powerful institution of markets fail in implementing Pareto efficient outcomes. This makes the design and analysis of appropriate institutions become a key issue for economic analysis and policy. The most prominent example for a situation where individual interests and the interest of the society as a whole are in conflict is, perhaps, the voluntary contribution to a public good, e.g., a clean environment. If the marginal cost of contributing to a clean environment exceeds the marginal utility of enjoying it, no incentive at the individual level to contribute to the public good exists. Nevertheless, it is of course well conceivable that every individual prefers a clean environment to a polluted environment, i.e., that a clean environment is the unique efficient outcome. In such a case, individual payoff maximization obviously leads to inefficient results.

*We thank Martijn Egas for valuable comments on an earlier version of this paper.

†University of Zurich, Institute for Empirical Research in Economics, Blümlisalpstrasse 10, 8006 Zurich, Switzerland

‡Tinbergen Institute and University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands

This article deals with such situations and possible solution mechanisms that might help overcome this fundamental conflict between individual incentives and social welfare. In fact, a whole series of conflicts arises in such "social dilemma" situations, the first being already at the individual level. An individual may recognize that cooperation (e.g., contribution to the public good) is overall beneficial. However, even if this is true, the individual also faces the inner conflict that unilateral cooperation basically means not to follow one's own material self-interest, i.e., behave differently from what would be optimal given any behavior of the other individuals. Nevertheless, suppose for the sake of the argument that some but perhaps not all individuals are able to resolve this inner conflict (e.g., through moral norms) and are willing to cooperate if and only if others cooperate as well. In this case, a second conflict arises, this time at the social level. As long as there are other individuals who free ride, i.e., do not cooperate, those, who in principal would be willing to cooperate, still do not cooperate. Because their behavior is conditional on the behavior of others, and there are others who do not cooperate, they also do not cooperate. A viscous circle which seems self-enforcing endlessly. However, what now if there existed a possibility to punish an individual's free-riding behavior? First, the threat of a new imposed conflict, namely the execution of punishment, might induce those individuals, who otherwise would never cooperate, to cooperate. Second, if successful, it will also induce those individuals who cooperate conditionally to cooperate as well. Thus, the threat of this third conflict might be able to resolve both the first (individual) conflict of those, who never cooperate, and the second (social) conflict of those, who are willing to cooperate but only conditionally. This article shall be about such resolution mechanisms.

We will discuss experimental evidence for two different institutional approaches to a possible resolution of this fundamental conflict between social welfare maximization and individual utility maximization. The basic workhorse for modeling this conflict is the voluntary contribution of a group of individuals to a public good. The common element of the two approaches is that both are based on the imposition of sanctions for free-riding behavior. The main difference between them concerns the question of "who punishes". In the first approach, punishment is executed by the group members themselves individually, i.e., punishment is decentral in nature. Each individual can decide, after contributions to the public good are made, whether he or she wants to punish other group members by reducing the

group members' payoffs. Of course, punishment may (and will) be costly to the punishing individual, as well. The crucial question is, whether individuals will punish at all, and if so, whether punishment will induce other individuals to cooperate. The second approach is based on the idea that individuals may be willing to delegate the punishment to a central, external authority. The classic example of such an institution is the constitutional state that maintains cooperation of its citizens through enforcement by central authorities (police, courts).[1] The essential element of such centralized sanctioning mechanisms is that, individual group members decide *ex ante* on the possible implementation of a centralized punishment institution, whereas in the decentralized sanctioning approach individuals can decide *ex post* on the possible execution of the punishment. The key questions to be answered are, whether individuals are willing and able to implement such punishment institution, how successfully implemented institutions look like, and whether they can increase the cooperation level of individuals in the group. We will address each of these question in the remainder of this article.

The remaining part is organized as follows. Section 2 presents a discussion of the decentralized punishment approach, then Section 3 analyzes the implementation of a centralized punishment authority. In both sections, theoretical foundations are given, but the focus will be on experimental evidence for both institutional approaches. Finally, Section 4 concludes.

2 Decentralized Punishment

In this section we discuss recent experimental evidence showing that cooperation can be successfully sustained beyond the well known folk-theorem results. In particular, we will present studies indicating that in situations where such classic mechanisms are most likely doomed to fail, people's disposition to sanction deviant behavior is a powerful device to curb selfishly motivated behavior. The sanction mechanisms presented are carried out individually and are therefore decentral in nature.

[1] Other examples include the EU Stability and Growth Pact created to enforce budgetary discipline of EU member states, or the Kyoto protocol aiming at a reduction of global CO_2 emissions by means of the implementation of legally binding agreements.

Fehr & Gächter (2000) first investigated the potential effect of individual punishment opportunities on voluntary contributions in a public good game.[2] The social dilemma that comes with a public good is linked to its properties of non-excludability and non-rivalry. Non-excludability means that once the good is provided nobody can be hindered from consumption and consequently benefiting from it. Non-rivalness implies that the consumption possibilities for the public good are not decreased by the number of people actually consuming it. Together, these two characteristics give rise to the well-known free-rider problem leading to inefficient underprovision of the public good and a conflict between selfish behavior and social welfare.[3]

2.1 The Voluntary Contribution Mechanism

A useful and commonly used workhorse for the experimental analysis of such social dilemma situations is the voluntary contribution mechanism. This mechanism models the tension between individual and common interest in its very basic and most clear-cut way. Consider a game with n players. Each player $i \in \{1, \ldots, n\}$ has an initial endowment ω and decides upon his or her contribution g_i to a joint project, the public good, with $0 \leq g_i \leq \omega$. Given the contribution decision of all players $g = (g_1, \ldots, g_n)$, the payoff to player i is given by

$$\pi_i(g) = \omega_i - g_i + a \sum_{j=1}^{n} g_j, \tag{1}$$

where $a < 1 < na$. The parameter a models the marginal per capita return from contributing to the public good. Since $a < 1$, a contribution of zero, i.e., $g_i = 0$, is the strictly dominant strategy for each player and hence $g = (0, \ldots, 0)$ is the unique Nash equilibrium of the game. Since $na > 1$, however, full contribution to the public good, i.e., $g_i = \omega$ for all i, is the welfare-maximizing strategy profile.

By now there exists already a large amount of experimental evidence for such public good games. Though some heterogeneity in behavioral patterns is observed

[2] In their seminal paper Ostrom, Walker & Gardner (1992) investigated the role and effectiveness of punishment in a common resource or tragedy of the commons problem. In such a setting a negative externality is imposed on others by the overuse of a common resource. In public good situations as discussed in this manuscript and investigated by Fehr & Gächter (2000) the externality imposed on others is positive by contributing to a project that benefits all.

[3] In the context of a common resource, like a pasture or fishing grounds, free-riding behavior leads to inefficient overconsumption of the common property, known as the "Tragedy of the Commons" (Hardin 1968).

the overall picture of subjects' behavior is similar across measurable traits like gender, and cultural and economic background. The basic picture emerging from this evidence is that a non-negligible amount of people exhibit a disposition towards voluntary contribution, at least when they are inexperienced with the game. When played repeatedly the typically observed contribution pattern is that on average contributions are rather high in the first rounds. Usually they mount to 40 to 60 percent of the initial endowment, depending on the value of a. With continuation of the game, however, voluntary contributions usually deteriorate and approach the selfish equilibrium of no contribution after a couple of rounds. Depending on the matching protocol there seems to be some differences in the dynamics of the contributions. In situations where people stay together in the same group and, hence, play the public good game repeatedly the contributions usually exhibit only a slight decrease, which is accompanied by a stark drop of voluntary contributions in the rounds close to the end of the game. In matching formats, where people also play the public good game repeatedly but with different group members in each round contributions often deteriorate more quickly and also show a less pronounced end game effect.[4]

Given the undesirable result of low or deteriorating voluntary contributions to a public good the question arises whether there exist mechanisms that sustain cooperative behavior in the long run. Fehr & Gächter (2000) examined precisely this question by investigating whether costly decentralized and private punishment can be effective in overcoming the free-rider problem.

2.2 The Effectiveness of Decentralized Sanctioning

To investigate the effectiveness of decentralized sanctioning, the public good game described above is appended by a second stage. At this stage all players in a group have the possibility to anonymously and individually punish other group members. After the first stage all group members are informed about the individual contributions to the public good, without revealing the identity of the respective members. Then, a group member i can punish another group member j by allocating punishment points p_{ij} to j. Each punishment point reduces the material payoff player j

[4]See Keser & Van Winden (2000) for a thorough experimental investigation into the differences in voluntary contributions in games with and without changing group members. For a general overview concerning behavior in public good experiments, see Ledyard (1995).

has earned on the first stage, the public good game, by 10 percent.[5] Importantly, the allocation of punishment points is not for free but comes at cost $c(p_{ij})$ for the punishing player i. The total material payoff Π_i of a player i who punishes other players j and also receives punishment points p_{ji} from these other players is given by

$$\Pi_i = \pi_i(g) \left[1 - \frac{\min(\sum_{j \neq i} p_{ji}, 10)}{10} \right] - \sum_{j \neq i} c(p_{ij}), \tag{2}$$

where $\pi_i(g)$ denotes the material payoff from the first stage (i.e., $\pi_i(g)$ comes from equation (1)).

Since the allocation of punishment points is costly, standard game theory - assuming common knowledge of rationality and purely selfish behavior - predicts that no player will allocate any punishment points in the second stage. By backwards induction, it follows that zero contribution in the public good game is still the dominant strategy. Hence, under the usual assumptions the equilibrium prediction is not altered by adding the punishment stage. There is, however, quite some evidence indicating that a non-negligible share of subjects exhibit traits that can be classified as reciprocal, in a broad sense. These are types of players who are ready to retaliate actions perceived as mean even if the retributive action is materially costly to themselves. If such types of players indeed exist and if they are actually ready do sacrifice material payoffs then the existence of punishment opportunities may change behavior considerably. In particular, it might mean that even players with selfish preferences may find it profitable to cooperate because they get punished otherwise.[6]

Fehr & Gächter (2000) tested this by implementing the above described two-stage game in the laboratory. Subjects participating in the experiment are divided into groups of four (i.e., $n = 4$) to play the game for several rounds. For control reasons a pure public good experiment without punishment opportunities is also conducted with the same set of participants. In some of the investigated treatments players stay together in the same group for all rounds (in the so-called partners

[5] To prevent losses there is an upper bound on the maximally possible reduction, which is 100 percent of the first stage earnings.

[6] Hauert, Nowak & Sigmund (2001) theoretically analyze this possibility for the case where players are endowed with binary strategies. In a completely different approach Fehr & Schmidt (1999) analyzed this game under the assumption that not all players are completely selfish but that at least some dislike to be worse and/or better off than other players in material terms.

protocol) whereas in others they are randomly re-matched with different players in each round. The parameters of the public good game used in the experiment are equal to $a = 0.4$ and $\omega = 20$. That is, the costs of punishment were increasing at an increasing rate. Importantly, however, the marginal cost to the punisher were strictly smaller than the marginal cost imposed on the punished person.

For the sake of comparison with the centralized punishment mechanism presented in the next section we concentrate on the partners protocol when reporting on the empirical findings. As mentioned above, Fehr & Gächter (2000) also conducted a control public good experiment without the punishment stage. This was combined with a punishment treatment in two variants. In one variant the public good game without punishment was played for ten rounds before the game with punishment opportunities was played also for ten rounds. In the second variant the order was reversed. That is, subjects played the public good game with punishment stage before the public good game without punishment possibilities. The results obtained for the game without punishment opportunities are in line with results from other public good games. In both variants the average contribution rates to the public good are initially rather high (around 45 percent when it was played first and slightly above 60 percent when it was played second) and far removed from the selfish prediction of zero contributions. Over time, however, voluntary contributions deteriorate, reaching rates of only about 15 percent in the last round, in both variants.

In the variant where the public good game with punishment opportunities was played first, at the beginning of the game the contributions are similar to those without punishment, with an average contribution rate of 55 percent. However, it turned out that the dynamics of voluntary contributions are starkly different, in the game with and without punishment. In the punishment treatment contributions steadily *increased* over time and reached 90 percent of the initial endowment in the last of the ten rounds. Also, in contrast to normal public good games, no end game effect was observed. A very similar pattern is found in the variant where the public good annex punishment game was played second. There contributions started at a rate of approximately 65 percent and increased up to more than 90 percent in the last two rounds.

These results show that the opportunity to punish seems to be a powerful tool to reduce free-riding behavior and facilitate contributions in the public good game.

The question remains whether the threat of being punished is already sufficient for promoting cooperation or if actual costly punishment is needed. It turns out that to some degree the bare possibility of being punished is already sufficient to discipline potential free-riders. This shows up in the higher average contribution in the first round of the public good game with punishment, compared to the first round of the game without punishment. However, this difference is only small and in many instances punishment is actually carried out, in particular at the beginning of the game. In these cases the picture that emerges is that punishment is the more severe the more the actual contribution of the punished person falls short of the average contribution from the rest of the group. If, for instance, a group member contributes between 14 and 20 points less than the rest of the group then this person receives between 6 and 7 punishment points from the fellow members. In contrast, if the deviation in contributions lies in the interval of -2 and $+2$ then the received punishment points are less than 1 on average.

A remaining question is whether the punishment opportunities are also welfare enhancing or if the costs of punishment outweigh the benefits from the higher contributions. It turns out that, on impact, punishment lowers overall payoffs. However, in the longer run free-riders learn that they get punished and refrain from such behavior. Therefore, actual punishment is not necessary anymore, implying that this decentralized punishment mechanism generates welfare gains, at least in the longer run.

In Fehr & Gächter (2002) the authors repeated the experiment with slight modifications with respect to the punishment technology. Instead of reducing the punished persons first-stage payoff by a percentage per point the subtraction takes place in absolute terms. Specifically, applying one punishment point inflicted a loss of three points on the punished person and created a cost of one point for the punishing person. More importantly, however, in this study the authors implemented a perfect strangers matching protocol. In such a matching procedure all players in a group meet only once and the participants know this. Compared to the above described partner protocol such a matching procedure has the advantage that reputation building is ruled out completely. Hence, any observed punishment must be non-strategic in nature, in the sense that it is impossible to build up a reputation of being tough. The behavior observed in this experiment is qualitatively and quantitatively similar to the one described above. Hence, even if reputation building is ruled out the

opportunity to punish seems to be a powerful device facilitating pro-social behavior in a social dilemma situation.

A number of other studies have shown the robustness of this result and also extended the results of Fehr & Gächter (2000) and Fehr & Gächter (2002). For instance, Carpenter (2003*b*) tested the robustness of the effectiveness of decentralized sanctioning with respect to the group size and the information players receive about the contributions of others. In his experiment, which is based on the design of Fehr & Gächter (2000), the author shows that increasing the group size from $n = 5$ to $n = 10$ rather increases than decreases contributions to the public good if decentralized punishment is possible. The likelihood to punish as well as the amount of punishment seems to be independent of group size. Information however seems to matter. The fraction of players whose contributions can be monitored and who can be punished has an impact on contributions. The more other players a single player can observe and (potentially) punish the higher are the contributions. In this study it seems that "information appears to be an important component of effective mutual monitoring" (Carpenter 2003*b*, p. 12) but that group size is not.

Anderson & Putterman (2003) and Carpenter (2003*a*) independently investigate whether the 'demand for punishment' in public good games follows the Law of Demand. The first authors run a 3-person public good experiment with a punishment stage, akin to the one in Fehr & Gächter (2002). Their matching protocol was of the perfect stranger format and the price of one punishment point was varied from 0 to 1.20. The main empirical findings of this work are that in all treatments contributions are relatively high and that they show almost no tendency to decrease over time. Punishment takes place and is mainly directed towards free-riders who contribute less than the average. Moreover, punishment *is* sensitive to price changes: lower prices lead to more punishment. Interestingly, punishment also takes place when the price of punishment is larger than the costs imposed on the deviant. The second author also changes the relative price of punishment exogenously. The relative price range investigated runs from 0.25 to 4 per punishment point. In this experiment players are in groups of size 4 and contributions of a player are revealed only to one other player who can then punish. The main findings are again that punishment is actually carried out and that it follows the Law of Demand. Furthermore, regression analysis indicates that punishment is a normal good and that it is inelastic with respect to income and price.

In a clever study Masclet, Noussair, Tucker & Villeval (2003) go one step further and ask whether non-monetary sanctions in the form of disapproval are already sufficient to induce pro-social behavior. For this they first replicate the Fehr & Gächter (2000) experiment with decentralized monetary punishment possibilities and extend it by also investigating the power of non-monetary informal sanctions. For that they gave participants the possibility to 'sanction' others by assigning 'disapproval points'. These points had no monetary consequences, neither for the person disapproved nor for the person disapproving. In the monetary punishment treatment the authors largely succeeded in replicating the data of Fehr & Gächter (2000). That is, the opportunity to punish induced higher contributions in the public good game and subjects were ready to punish those perceived as free-riders. Interestingly, at least on impact the non-monetary sanctioning mechanism was as effective as the monetary punishment mechanism. Compared to a control treatment without any opportunity to sanction the possibility of verbal disapproval lead to significantly higher contributions to the public good. Over time, however, monetary punishment turned out to be more effective than non-monetary sanctioning. It is, therefore, an open question whether the effectiveness of non-monetary sanctioning survives in the long run.

Given the surprising effectiveness of decentralized sanctioning naturally the question arises whether the same effect can be attained with the help of decentralized *rewarding*. Sefton, Shupp & Walker (2002) investigated this possibility by running three different sets of experiments. All three treatments consisted of two sequences. In the first sequence a public good game without a second stage was played by experimental subjects in groups of four for 10 rounds ($\omega_i = 6, a = 0.5$). Thereafter, the game was played for another 10 rounds with a second stage in each round. This second stage offered the possibility to either sanction other players by costly taking away some money, or reward other players by costly giving some money, or to do either of the two, depending on the treatment. In the sanctioning case the costs were the same for both players, the punished and the punishing one. When rewarding the cost for the rewarding player equaled the benefit of the rewarded player.[7] The

[7]Compared to the set-up used by Fehr & Gächter (2000) there is one potentially important difference in the design of the second stage. Whereas in Fehr & Gächter (2000) punishment had to be paid out of the earnings of the first stage, in Sefton, Shupp & Walker (2002) subjects received an extra endowment of $\omega_i = 6$ that could be used for punishing and/or rewarding.

results of the treatment with sanctioning are qualitatively similar to those reported by other studies, though quantitatively the contributions fall behind those reported in Fehr & Gächter (2000). Interestingly, it turns out that rewarding is less effective in promoting cooperation. This holds, in particular, in the longer run. Consequently, in the treatment with the possibility of rewarding and sanctioning subjects tend to sanction free-riders more often and heavier than they reward 'above-average contributors'. A plausible reason for this result is that the threat of sanctioning is already sufficient to induce potential free-riders refraining from doing so. Rewards, in contrast, have to be actually paid out to be effective.

3 Centralized Punishment

The experiments discussed in the preceding section show that individuals are able to sustain substantial degrees of cooperation in repeated public goods games. It is important to stress that this result goes well beyond the classic folk theorem result: cooperation is sustained even if there is no rational and selfish incentive in the institutional environment for an individual to cooperate. The main mechanism sustaining cooperation is the willingness of individual group members to sanction others' free-riding behavior, even if this is costly to themselves.

Private, decentralized punishment, however, may not be possible in every situation. For example, group members may not have enough detailed information about the individual contributions of others but only average data about the players' behavior. This may be the case, e.g., when groups are very large or when, spatially, individuals are located far apart from each other. In other cases, no information may be available at all. Obviously, under such circumstances private sanctions do not make much sense since defectors can not be identified. As discussed before, Carpenter (2003*b*) also finds support for the importance of information with regard to private punishment.

Even if individuals are informed about the contribution decisions of other group members, it may still be the case that an appropriate punishment technology is not available to individuals in the group. For example, the reduction of one's own contribution, a strategy that in principle is always available, is not an appropriate punishment technology, since every group member is punished in the same way independent on whether he contributed to the public good or not. This form of pun-

ishment does obviously not allow for the targeting of individual defectors. Finally, even if a perfect punishment technology of the type implemented in the Fehr-Gächter experiments is available, it may still be too costly for individuals to be used. The results of Anderson & Putterman (2003) and Carpenter (2003*a*) show that a subject's willingness to punish other group members obeys the classic law of demand. That is, the impulse to punish declines as the price of punishment increases.

If private sanctions are too costly or the mechanism of decentralized punishment is not available at all, a possible alternative may be to delegate punishment to an external enforcement agent, i.e., to create a "Leviathan" that watches individuals' behavior and punishes in case of defection. Looking at the history of our society we clearly see many institutional solutions of this form. The main difference to the private punishment institutions discussed before is that now, private and voluntary sanctions are substituted by a centrally organized punishment system. Similarly, as the crucial question was before whether individuals are willing to punish other group members, the question is now whether individuals can voluntarily create a centrally organized institution that takes care of the punishment and thereby sustains cooperation. For example, do all group members participate in such institution formation or do some individuals try to free-ride on the formation process? What implication does the existence of free-riders have on the creation of an institution? In this section we want to address these question both theoretically and empirically.

3.1 A Game-theoretic Model of the Leviathan

Okada (1993) and Okada (1997) proposes a non-cooperative game model to study the voluntary implementation of a centralized punishment institution in case of a n-player prisoners' dilemma game. We can easily extend his analysis to public good games. Consider an n-player public good game as described in the preceding section. There are n players, each with an initial endowment ω who individually decide how much to contribute to a public good. Individual payoffs are exactly as defined before; given the contribution decision of all players $g = (g_1, \ldots, g_n)$, the payoff to player i is given by $\pi_i(g) = \omega_i - g_i + a \sum_{j=1}^{n} g_j$, with marginal per capita return a satisfying the condition $a < 1 < na$.

The main idea of Okada's model is that, before players come to play the above public good game they can vote on forming an institution that has the sole right

to punish any player who contributes less than his full endowment. Thus, different than before players do not have the possibility to punish *ex post*, but can *ex ante* create a centralized agency that controls contributions and exerts the punishment. The process of institution formation is modeled by the following three stage game.

In stage one, the so-called *participation stage*, each player i announces whether he is willing to participate in such an institutional arrangement, where he is punished in case he does not contribute his full endowment to the public good. It is assumed that the punishment institution is costly and that total costs are equal to $c > 0$. Let $S \subseteq \{1, \ldots, n\}$ be the set of players who are wiling to participate in such a punishment institution. All players are informed about the outcome of stage one, i.e., about the set S.[8]

In the second stage, the *negotiation stage*, members of S vote on whether the punishment institution shall be implemented for all players in S. More generally, the negotiation stage may also include a collective bargaining process, where members of S decide upon the effective punishment level and the sharing rule for the institutional cost c. Let us abstract from this bargaining process, however, and analyze the reduced form of the game by assuming that institutional punishment is large enough to induce full contribution to the public good and that costs are shared equally among all members of S.[9] The institution is implemented if and only if all players in S vote for it, i.e., an unanimity rule is used. Note that a player who is not a member of S cannot vote for the implementation of the institution, and that the institution will only be able to punish members of S.

Finally in stage three, the *contribution stage*, the public good game is played, where depending on the decisions in the preceding stages two things can happen. If $S \neq \emptyset$ and all players $i \in S$ vote for the institution in the negotiation stage, each player $i \in S$ contributes his full endowment to the public good and pays his part of the total cost equal to $c/|S|$. At the same time each player $i \notin S$ freely decides on his contribution g_i and also pays no cost. On the other hand, if either $S = \emptyset$ or at least one member of S votes against the implementation of the punishment institution in stage two, no punishment is implemented and all players play the original public good game.

[8] Alternatively, it is also possible that players are only informed about the cardinality of S (see below).

[9] See Okada (1993, 1997) for a more general analysis of the bargaining process.

Formally, the resulting payoffs in the modified game are as follows. If the punishment institution is implemented, for every player $i \in S : g_i = \omega$ and hence

$$\tilde{\pi}_i(g) = a \sum_{j \in S} \omega + a \sum_{j \notin S} g_j - \frac{c}{|S|} \tag{3}$$

if $i \in S$, and

$$\tilde{\pi}_i(g) = \omega - g_i + a \sum_{j \in S} \omega + a \sum_{j \notin S} g_j \tag{4}$$

if $i \notin S$. If the punishment institution is not implemented, each player is free to decide how much to contribute, i.e., payoffs are given by equation (1) for all i.

The following proposition is a result from Okada (1997), which applies backward induction to characterize subgame perfect Nash equilibrium.

Proposition 1 *A punishment institution is implemented in a strict Nash equilibrium in the negotiation stage if and only if the number of players participating is at least s^*, where the latter is the minimum integer s satisfying the condition*

$$sa\omega - \frac{c}{s} > \omega. \tag{5}$$

In consequence, an action profile in the participation stage is a strict Nash equilibrium iff exactly s^* players participate.

The intuition for Proposition 1 is as follows: without punishment institution the unique Nash equilibrium in the contribution stage is the strategy profile where all players contribute nothing, thereby earning a payoff of ω each. In consequence, in the negotiation stage a punishment institution is implemented if and only if the payoff from forming that institution and contributing to the public good is larger than ω. This is expressed by condition (5), which is satisfied if at least s^* players join the institution. Since it is always a best response for a player not to join the institution and to defect if $s \geq s^* + 1$ players already join (because the other players still have a strictly positive incentive to form the institution and the punishment institution does not apply to players who are not a member of it), no more than s^* players will implement the punishment institution in the unique strict Nash equilibrium of the game.

Note that this equilibrium prediction depends on two crucial behavioral assumptions: first, no player is willing to contribute to the public good if there is no

punishment institution and second, every player cares only about his own material payoff. In particular, the proposition says that players will form an institution and contribute to the public good if it individually pays to do so, independent on whether other players decide not to join and to defect, thereby fully exploiting the public-good benefits created by the punishment institution. That there exists a substantial fraction of individuals who are willing to contribute to a public good even without a threat of punishment, is by now an empirical fact. However, whether individuals care about the behavior of others, especially about possible free-riders, when creating a centralized punishment institution, is an open question. We can address this question by designing an appropriate laboratory experiment.

3.2 The Implementation of Centralized Punishment

The experimental design basically follows the game-theoretic protocol of Okada's model. Subjects invited to the laboratory are divided into groups of size n, who play the above described three-stage game. Since our main interest is on subjects' behavior with regard to the implementation of the institution and not on the well-functioning of the institution itself (e.g., the effective execution of punishment, or the behavioral consequences of centralized punishment), we can modify the game slightly and assume that subjects who have implemented the centralized punishment institution, do no longer make a decision in stage three but are forced to contribute their whole endowment to the public good. The idea is that, by forming an effective punishment institution, subjects simply bind themselves to full contribution.

With this assumption, each subject decides in stage one whether he is willing to bind himself to contribute his full endowment. Given this decision, in stage two all subjects learn how many of the other subjects are willing to bind themselves and next, those subjects, who are willing to bind themselves, vote on whether the institution shall be implemented or not. If all subjects eligible to vote agree on the implementation, in the final stage their contribution is set equal to their whole endowment while all others decide how much to contribute. In addition, each subject who is bound pays a cost equal to the per capita cost of the institution. If at least one subject eligible to vote does not agree on the implementation, in the final stage all subjects play the original public good game. That is, no subject is bound to full contribution, every player freely decides how much to contribute, and no institutional

costs arise.

Kosfeld, Okada & Riedl (in preperation) have conducted the above described experiment with parameters of the public goods game equal to $n = 4$, $\omega = 20$, and $a = 0.65$. They consider various treatments. In one treatment the total cost of the institution is set equal to $c = 2$, which — using Proposition 1 — leads to the prediction that exactly $s^* = 2$ subjects implement the institution and contribute their full endowment, whereas the remaining two subjects free-ride and contribute zero points to the public good. Thus, in equilibrium subjects who are bound to full contribution earn a payoff of $0.65 * 40 - 1 = 25$, while a subject who is not bound and does not contribute earns a payoff of $20 + 0.65 * 40 = 46$. Subjects play 20 periods of this game in the experiment with the composition of groups being constant in every period. Thus, the matching of subjects in the experiment is the same as in the partners protocol in Fehr & Gächter (2000).[10] All results are based on the independent observations of 11 different groups.

In what follows we will focus on the following empirical questions: (1) Are subjects willing to bind themselves, or are institutions based on the voluntary self-control of its members unlikely to be observed? (2) Do we see the institutions that are predicted by the theory, i.e., are most institutions of size $s^* = 2$, or do subjects form other institutions? (3) Does the formation of institutions have a positive effect on contribution rates, i.e., do subjects contribute more to the public good if they have the possibility to bind themselves than if they do not have this possibility?

Question (1) can be answered in a positive way. in 132 out of 220 times (60 percent) do subjects implement a centralized punishment institution that binds at least one subject to contribute his full endowment to the public good. Note that if subjects played according to the prediction of Proposition 1, all groups would form an institution in every period, which is clearly not the case. Moreover, the size of the institution would always be equal to two. However, as the following result shows this is not the case.

The large majority (68 percent) of the institutions observed in the experiment are of full size, i.e., $|S| = 4$. Only 12 percent of the institutions are of size two, 16 percent

[10] The repetition of the game is important because subjects have to learn a lot in the experiment and it is very unlikely that groups "jump" into equilibrium already in the first period. Furthermore, since most real-life public good situations that allow for the implementation of a centralized punishment institution involve the repeated interaction of the same group members, a partner design seems also appropriate.

are of size three, and 4 percent are of size one. Thus, two out of three institutions contain all group members. Institutions formed by less members are rarely seen. This result is not driven by the fact that subjects almost always propose full-size institutions, i.e., that all subjects declare their willingness to participate in the first stage. Quite to the contrary, in 55 percent of the cases in stage one at least one subject declares not to be willing to participate. However, such institutions where less than four subjects participate have a very low chance to be implemented. While the probability to reach unanimous agreement in the negotiation stage is about 0.91 if all members are willing to participate, the probability falls below 0.38 if there exist at least one group member who does not want to participate.

Given that the implementation of an institution fails in about 40 percent of the cases, an important question is whether overall, the possibility to form such institutions is a good thing, i.e., whether on average the level of contribution to the public good is raised. To answer this question, contribution rates are compared to those in a control treatment, where subjects play 20 periods of the same public goods game without the possibility of institution formation. The authors find that in the control treatment average contribution to the public good is 13.5 points during period 1 to 10 and falls to 11.5 points during 11 to 20. In contrast, if subjects can form a punishment institution, average contribution is 13.6 points in the first half of the experiment and increases to 14.6 points in the second half of the experiment. Thus, even if the process of institution formation may fail, contribution levels are raised on average at least in the later rounds.

To summarize, the experimental findings of Kosfeld, Okada & Riedl (in preperation), while still preliminary at this stage, clearly indicate that the voluntary implementation of a centralized punishment institution may represent an important mechanism for solving the free-rider problem in social dilemma games. Although the process of implementation is rather complex and therefore it is not unlikely that the implementation fails, contribution levels are raised on average in the experiment. Moreover, if the implementation is successful, in most cases the welfare maximizing outcome is realized: all group members participate and each member contributes his full endowment. This observation is particularly interesting as the theoretically predicted institution size is much smaller, i.e., two instead of four. Thus, total contribution in the experiment under a centralized punishment institution exceeds predicted total contribution by almost 100 percent.

4 Conclusions

The design of effective mechanisms to overcome the free-rider problem in public good provision and common resource problems is a recurring theme in economics and the political sciences. That institutions are important for solving problems like over-fishing and the greenhouse effect is a truism at least since the seminal works of Hardin (1968) and North (1990). However, till recently clear evidence on the effectiveness of different institutional arrangements designed to overcome social dilemma situations was missing. In particular, it was largely unknown what institutions are effective not only theoretically but also empirically.[11]

In this contribution we present and review recent experimental studies on the effectiveness of sanctions as a cooperation device. We concentrate on two approaches tackling the social dilemma problem from two different angles: voluntary decentralized and centralized punishment. The experimental literature on voluntary decentralized sanctioning started with Ostrom, Walker & Gardner (1992) and was picked up again and further investigated by Fehr & Gächter (2000). Both studies show that decentralized punishment can be very effective in overcoming the free-rider problem. Since then a stream of experiments was conducted showing the robustness of this mechanism. [12]

Concerning the effectiveness of voluntary centralized sanctioning mechanisms much less is known. Kosfeld, Okada & Riedl (in preperation) investigate whether experimental subjects are ready and able to costly found a centralized sanction institution. They examine the influence of the costs of the institution and its effectiveness relative to a world without any form of sanctioning opportunities. In the investigated institutional setting it is shown that a centralized institution is formed if costs are low and that institutions (almost) only form when all group members are joining. This holds despite the theoretical prediction of institution forming at any cost and the theoretical possibility of institutions where not all people join.

[11] This seems to be at least partly due to the fact that many of the theoretically developed contribution mechanisms are rather complex. For an exception see Falkinger, Fehr, Gächter & Winter-Ebmer (2000).

[12] Other less direct decentralized punishment possibilities are also shown to be effective. Riedl & Ule (2004) show that the possibility to individually exclude others from one's network of interaction can have dramatically positive effects on the cooperation rates in social dilemma situations. Partner selection (Hayashi & Yamagishi 1998, Coricelli, Fehr & Fellner 2004) and ostracism (Masclet no date) seem also to be strong forces disciplining free-riders and fostering cooperative behavior.

What stands out - next to the effectiveness of the decentralized punishment institutions - is the fact that theoretical models based on common knowledge of rationality and egoistic preferences largely fail to predict actual behavior. Motivations of reciprocation, retaliation, and other-regarding preferences seem to play an important role in shaping people's behavior, in particular in social dilemma situations. It seems important to take this behavioral dispositions into account also when designing centralized sanction institutions. Otherwise, as in the experiment of Kosfeld, Okada & Riedl (in preperation), people may refrain from participating in such institutions thereby foregoing potentially large efficiency gains. This requires much effort for the development of new and better models for designing institutions for cooperation. Also empirically much work has still to be done till we may have a good understanding of the principal driving forces of cooperation within particular institutional structures.

References

Anderson, Christopher M. & Louis Putterman. 2003. "Do Non-strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism." mimeo.

Carpenter, Jeffrey P. 2003*a*. "The Demand for Punishment." mimeo.

Carpenter, Jeffrey P. 2003*b*. "Punishing Free-Riders: how group size affects mutual monitoring and the provision of public goods." mimeo.

Coricelli, Giorgio, Dietmar Fehr & Gerlinde Fellner. 2004. "Partner Selection in Public Goods Experiments." Working paper, Max Planck Institute, Jena.

Falkinger, Josef, Ernst Fehr, Simon Gächter & Rudolf Winter-Ebmer. 2000. "A simple mechanism for the efficient provision of public goods - experimental evidence." *American Economic Review* 90:247–264.

Fehr, Ernst & Klaus Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114:817–868.

Fehr, Ernst & Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90:980–994.

Fehr, Ernst & Simon Gächter. 2002. "Altruistic punishment in humans." *Nature* 415:980–994.

Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162:1243–1248.

Hauert, Christoph, Martin A. Nowak & Karl Sigmund. 2001. "Reward and punishment in minigames." *Proceedings of the National Academy of Sciencies USA* 98:10757–10762.

Hayashi, Nahoko & Toshio Yamagishi. 1998. "Selective play: Choosing partners in an uncertain world." *Personality and Social Psychology Review* 2:276–289.

Keser, Claudia & Frans Van Winden. 2000. "Conditional cooperation and voluntary conmtributions to public goods." *Scandinavian Journal of Economics* 102:23–39.

Kosfeld, Michael, Akira Okada & Arno Riedl. in preperation. "Institution and Cooperation." Tinbergen Institute and Institute for empirical economic research.

Ledyard, John O. 1995. Public Goods: A Survey of Experimental Research. In *The Handbook of Experimental Economics*, ed. John H. Kagel & Alvin E. Roth. Princeton: Princeton University Press pp. 111–194.

Masclet, David. no date. "Ostracism Applied to a Public Good Game." mimeo.

Masclet, David, Charles Noussair, Steven Tucker & Marie-Claire Villeval. 2003. "Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism." *American Economic Review* 93:366–380.

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge, UK: Cambridge University Press.

Okada, Akira. 1993. "The Possibility of Cooperation in an n-Person Prisoners' Dilemma with Institutional Arrangements." *Public Choice* 77:629–656.

Okada, Akira. 1997. The Organization of Social Cooperation: A Noncooperative Approach. In *Understanding Strategic Interaction, Essays in Honor of Reinhard Selten*, ed. W. Albers et al. Springer-Verlag pp. 228–242.

Ostrom, Elinor, James Walker & Roy Gardner. 1992. "Covenants with and without a sword: Self-governance is possible." *American Political Science Review* 86:404–417.

Riedl, Arno & Aljaž Ule. 2004. "Cooperation, Exclusion, and Social Structure in Network Experiments." Tinbergen Institute.

Sefton, Martin, Robert Shupp & James Walker. 2002. "The Effect of Rewards and Sanctions in Provision of Public Goods." Working Paper 2002-2, CeDEx, University of Nottingham.