Ubiquity and efficiency of restrictions on informal punishment rights

KIMMO ERIKSSON[1,2], PONTUS STRIMLING[1], MICAEL EHN[1]
[1]Centre for Study of Cultural Evolution, Stockholm University, Sweden

[2]*School of Education, Culture and Communication, Mälardalen University, Västerås, Sweden*

Abstract

Over-punishment often occurs in anonymous peer-to-peer punishment in public goods game experiments where punishment is free for all. We report a public goods game experiment in which a condition where punishment rights were restricted to one other player per player yielded higher total welfare than a condition with unrestricted punishment. In the restricted punishment condition, there was much less punishment but high levels of cooperation were achieved nonetheless. This indicates that it may be beneficial to groups to restrict punishment rights. In a second study we presented respondents from many different countries with three scenarios constituting everyday social dilemmas of various kinds. Across countries, respondents tended to judge it as inappropriate for most involved parties to punish selfish individuals in the scenarios. Typically, only one party was judged to have the right to punish. Whereas much prior work has considered punishment as a public good that needs to be encouraged, these findings suggest that informal norms about sanctions tend to constrain punishment to certain individuals. Such norms may serve the function to harness the positive effects of punishment while containing the negative effects, and we suggest that they are likely to arise from learning.

Keywords. cooperation; punishment; rewards; social norms; right to punish

This is a postprint version of: Eriksson, K., Strimling, P., & Ehn, M. (2013). Ubiquity and efficiency of restrictions on informal punishment rights. *Journal of Evolutionary Psychology, 11*(1), 17-34. https://doi.org/10.1556/JEP.11.2013.1.3

Introduction

The topic of punishments is controversial. Whereas some researchers promote peer-to-peer punishment as key to human cooperation (e.g., Fehr & Gächter, 2002), others argue that this is actually an unsuccessful strategy (e.g., Dreber et al., 2009). These positions reflect that punishment has both an upside and a downside. The upside is potential deterrence of unwanted behavior. The downside is more multifaceted. Various psychological mechanisms may make threats of punishment ineffective as a deterrent (Netter, 2005). There are also costs of various kinds, in addition to the cost inflicted on the punishee: there is the punisher's effort, the punishment's potential damage to future relations, and costs that arise because punishment is sometimes used in ways that are not conducive to cooperation. Accordingly, several laboratory studies show that the net effect of punishment is not necessarily positive (e.g., Herrmann et al. 2008; Nikiforakis, 2008; Nikiforakis & Normann, 2008). Here we are interested in how social norms may deal with this problem by restricting punishment so that its potential as a force for increased welfare is harnessed.

First note that a related but complementary solution is to formally place sanctioning powers in the hands of a central agency, e.g., a law-enforcing police (Baldassarri & Grossman, 2011). As famously observed by Max Weber, a central agency that adopts the responsibility to punish certain behaviors by coercion typically claims *Gewaltmonopol*, i.e., prohibits peer-to-peer coercion. A single central agency could, at least in theory, set punishment at a level that optimizes welfare. There is some experimental evidence that groups are better off when punishment is delivered by a central agency than when group members sanction each other (O'Gorman et al., 2009).

Here we are not concerned with central agencies but instead turn our attention to those situations where decentralized sanctions, often called informal sanctions, are actually used. Such situations occur when anarchy makes informal sanctions necessary (Amster 2003), but, more importantly, non-violent informal sanctions are often used when the offense is not on the level recognized by the central agency. Even in these situations, we claim that individuals are not free to punish each other at will, because strong social norms

regulate who can informally sanction whom. Our second study below is an international survey designed to demonstrate this claim.

A complementary way of restricting punishment is to impose limits on the severity of punishment. Such norms have been identified as one of the key features of durable informal institutions for management of common pool resources (Ostrom, 1990). In brief, restrictions of the *severity* of punishment are well-known and have been the subject of influential research. Our focus here is instead on norms about *who* has no right to sanction another. Specifically, we propose that in a given situation with many potential punishers it may be illegitimate for most individuals to actually punish. Further, this illegitimacy is not inherent to the individual but based on context-specific roles. Research conducted in France has shown that both potential punishers and potential punishees perceive differences in roles as extremely important for who can legitimately sanction deviant behavior (Nugier et al., 2007; Chaurand & Brauer, 2008). Remarkably, this important notion that most agents may be normatively *discouraged* to punish selfish behavior seems to have been completely neglected in the large body of research on cooperation in social dilemmas.

This neglect of norms against punishment may have its roots in the dominance of a game theoretic perspective on social dilemmas. Game theory assumes that agents are rational. In a social dilemma, the individually rational strategy is to defect. This problem can potentially be solved by a threat of sanctions against defection – but if it is somehow costly to sanction others, a second-order free-rider problem arises where the individually rational strategy is to shirk from sanctioning. Thus, in order to uphold norms against defection, it is necessary to also uphold norms against shirking from sanction of defectors. Such norms to encourage punishment were called "metanorms" by Axelrod (1986), an early proponent of the game theoretic approach.

However, experiments on behavior in games have repeatedly demonstrated that the rationality assumption of game theory is not always valid. In particular, people do not seem to use punishment as game theory would predict. For instance, it has been demonstrated again and again that many people do use costly punishment even when they are not sanctioned for not doing so (e.g., Yamagishi, 1986; Fehr & Gächter, 2002; Herrmann et al., 2008; Nikiforakis & Normann, 2008). It has also been shown that if a second stage

of punishment is available, it is more often used against those who punished than against those who did not punish in the first stage (Cinyabuguma, Page & Putterman, 2006). It is not difficult to find emotional underpinnings of these instances of "irrational" behavior. In real life, people get angry at each other and may act on that anger whether it is rational or not (Lerner & Tiedens 2006). Specifically, people may get angry at someone they think behaved selfishly, and they may derive satisfaction from punishment of this individual (de Quervain et al., 2003; Singer et al. 2006). Because punishers tend to be seen as *not* taking others' interests into account (Strimling & Eriksson, forthcoming), they may draw punishment from others who get angry at them. From an evolutionary viewpoint, there is of course nothing mysterious about anger sometimes producing irrational behavior. It is sufficient that anger has been adaptive on average. Even if using costly punishment in certain lab experiments is not to the individual's advantage, it easy to conceive of many situations in which it is beneficial for the individual to have a capacity for costly aggressive protection of his interests. What we are interested in here are the problems for the group that a propensity to punish others creates and the social norms that therefore should restrict people from using punishment.

In the beginning we mentioned the multifaceted downside of punishment. In this paper we shall focus on the aspect of over-punishment. As discussed above, experiments show that many individuals are willing to punish others in social dilemmas. When someone behaves in a way that many others are willing to punish, and punishment is free for all, the same act may be punished several times. This is likely to sum up to more punishment than is necessary for deterrence. Let us quickly sketch what a formalization of this argument could look like: Assume that many agents are conditional cooperators and altruistic punishers, i.e., they will initially make high contributions and they will punish those who make low contributions in proportion to how low they are. Assume that other agents are rational, so that they would match the higher contributions of others if and only if they otherwise expect to be punished more than they would gain from making a low contribution. If every agent has just a single potential punisher, a rational agent will match the expected higher contribution of the punisher if they receive punishment at least as great as the proportion b of a contribution that is lost to the contributor. In typical experiments, b equals one half and punishment is just a third as costly to the punisher as to the punishee, in which case a single potential punisher is sufficiently

deterrent if he can be expected to pay at least one unit to punish a low contributor for every *six* units the low contribution is lower than his own. Further, if deterrence of low contributions is thus achieved with a single punisher, then having more than one punisher per punishee would be unnecessary. Of course, deterrence is not achieved immediately – agents would first need a learning period where they develop expectations of each other's behavior. During that learning period punishment would sometimes be used. When punishment is used, it is done at a higher total cost if there are more punishers per punishee. This extra cost can then be regarded as over-punishment.

The model sketched above is designed only to capture the idea that if many people are voluntary punishers, free for all punishment may lead to more punishment than is necessary. The model of course ignores many real features of the problem. It is therefore very important to see whether experimental restrictions of free for all punishment will actually lead to increased welfare. One previous study supports this conclusion: O'Gorman et al. (2009) argued that avoiding over-punishment was one potential benefit of centralization of punishment and they demonstrated in a public goods experiment that a condition where punishment was restricted to a single punisher yielded higher mean profits than a condition where everyone could punish. Our argument above suggests that centralization to a single punisher is just one special case of restricting punishment rights and that the problem of over-punishment may be addressed also by non-centralized restriction of punishment rights. Our first study below is an experimental approach to this issue. Specifically, this study is aimed at demonstrating that also non-centrally restricted peer-to-peer punishment yields higher welfare than unrestricted punishment.

The second question is to what extent punishment rights are, in fact, restricted. As we mentioned in the beginning, central authorities tend to restrict individuals' right to punish each other using violence. Our proposition is that even for the kind of punishment that is allowed between individuals, such as humiliating or yelling at the punishee, social norms informally restrict punishment rights to a considerable extent. As a consequence, situations regulated by central authorities and situations regulated by social norms among peers will not be so different after all. In particular, we propose that social norms restrict punishment rights in everyday social dilemmas. In other words, we propose that even in situations where several people suffer

from one individual's selfish behavior, most of the involved parties tend to be *discouraged* from punishing the selfish individual. We investigate this proposition in an international web survey (Study 2).

Above we discussed norms that limit over-punishment as beneficial to the group. We do not, however, suggest that such norms arise from some kind of group selection process. Indeed, selection can never explain why something arises, only why it spreads. For something as complex as social norms, the first priority must be to understand the process whereby they arise in the first place. Only then can it be evaluated whether subsequent selection is necessary to explain the popularity of a certain norm, or whether the initial process whereby norms arise might be sufficient to explain also the popularity of certain norms. Although an investigation into this is outside the scope of this paper, in the discussion we sketch a process whereby norms that restrict punishment can be expected to arise through individual learning within a cultural context. We strongly encourage research on how norms in general, and norms about use of punishment in particular, arise.

Study 1

The objective of the first study was to demonstrate that, by avoiding over-punishment, even non-centralized restrictions on punishment rights can increase the efficiency of cooperation is a social dilemma. Non-centralized restrictions can be structured in many different ways, and in real life we expect many different structures to appear depending on the context (see Study 2). In this experiment we simply chose a structure maximally different to a centralized structure. Centralization of punishment amounts to a single punisher who can punish everyone, with everyone else restricted from punishing at all, as in the study of O'Gorman et al. (2009). To achieve maximal non-centralization, we used a structure where every group member was restricted to punish exactly one other player, namely, the next player along a cycle (Figure 1).[i]

[i] This cyclic punishment structure has been used in one previous study (Carpenter 2007). However, that study also restricted information on contributions such that every player was informed only of the contribution of his or her designated punishee. In contrast, we are here interested specifically in the effect of restricting punishment rights, holding other things constant.

Our prediction is that mean profits will be higher if punishment is restricted in this non-centralized way than if punishment is unrestricted.

Participants

One hundred and sixty participants (55% male, average age 24 years) were recruited from a pool of people who had previously registered as interested in taking part in game experiments. Participants were mainly students all across Stockholm University. Twenty sessions took place with 12 participants per session. The part of the session devoted to this experiment lasted approximately half an hour. Each monetary unit (MU) earned during the session equated to SEK 0.15.

Design and procedure

Participants were randomly divided into groups of size four. In order to model everyday situations where the same groups interact repeatedly (as in Study 2), groups were fixed throughout the session and group members could keep track of the other three group members through persistent identifiers (the numbers 1 through 3). Keeping these features constant, the experiment aimed at investigating the difference between restricted and unrestricted punishment rights, as detailed below.

In each round of the game, every participant received an endowment of 20 MUs and decided how much of this endowment to contribute to the common pot. Participants were informed that contributions to the common pot were doubled and distributed equally to all group members. Thus each MU invested in the common pot yielded a payoff of 0.5MU to each group member, irrespective of who invested, as in many other studies (e.g., O'Gorman et al., 2009). After contributions were made, every player was informed of each group member's contribution. See Appendix 1 for complete instructions. Groups played eight rounds in a no-punishment treatment followed by eight rounds in a punishment treatment. The punishment treatment existed in two different conditions, with 20 groups in each condition:

(1) *Unrestricted punishment:* Every participant was asked for each of the other group members, after being informed of their respective contributions, how many MUs he or she wanted to use to reduce the

payoff of that group member. For each MU paid to punish someone, three MUs were deducted from the punishee's earnings.

(2) *Restricted punishment:* Same as the previous condition except that participants could only punish their designated punishee. As depicted in Figure 1, punishees were assigned cyclically (and this assignment was fixed).

The total profit of a group over a treatment was measured as the sum of all payoffs to group members over the eight rounds of the treatment, from which sum the total endowments (8 rounds × 4 group members × 20 MUs per round and group member = 640 MUs) were subtracted. In the no-punishment treatment, the total profit simply equaled the total contribution to the common pot, because every MU contributed to the pot created an extra MU of total payoff. In the punishment treatment, the total profit equaled the total contribution minus the total cost of punishment, which is 4 times the total number of MUs paid to punish others.

Results

Table 1 presents descriptive statistics of how total contributions, total punishment costs and total profits depended on treatment and condition. Figure 2 shows per round profit. As expected from random assignment of conditions, there was no significant difference between conditions in the no-punishment treatment, t_{38}=0.75, p=0.46. In the punishment treatment, total profits were higher when punishment was restricted than when it was unrestricted, t_{38}=2.42, p=0.02, d=0.72. This difference was driven by total punishment costs being lower when punishment was restricted, t_{38}=3.39, p=0.002, d=0.95; there was no significant difference in total contributions, t_{38}=0.53, p=.60.

Discussion

The results of this experiment clearly support the notion that non-centralized restriction of punishment rights can increase the efficiency of cooperation in a public goods game by reducing over-punishment. It should be noted that in the last few rounds of the experiment profits increased in the unrestricted punishment condition (see Figure 2), indicating that the problem of over-punishment diminished over time. One reason

behind this was a tendency for punishees to be punished by fewer punishers later in the game; the Spearman correlation between the average number of punishers of those who were punished at all and the round of the game was negative: rho(8)=-.69, p=.058. This can be interpreted as spontaneously developing role-taking, which is one way in which norms about who can and who cannot punish who could develop. We return to this topic in the general discussion.

Study 2

The objective of the second study was to demonstrate the ubiquity of informal restriction of punishment rights in everyday social dilemmas. We used three scenarios chosen to represent three basic types of social dilemmas: depletion of a common resource, free-riding on a joint effort, and pollution of a common environment. In each scenario, there were several parties in different roles who all suffered from someone's selfish behavior. Our prediction was that most people nonetheless would find it inappropriate for some of these parties to punish the selfish behavior, restricting punishment rights to only a single role in each scenario.

Participants

An online survey was completed by 528 participants (63% male; mean age 30 years) of mixed educational backgrounds and from 73 different countries. Participants were recruited using the Amazon Mechanical Turk (https://www.mturk.com) and received a compensation of half a US dollar.

Design and procedure

The questionnaire presented three scenarios (see Appendix 2). In brief, the scenarios described (a) two families dining together and discovering that one of the children has already eaten the sweets meant for dessert; (b) a hospital ward where one nurse has come in to work very late such that the others have been forced to work extra hard; (c) a student apartment where one of several roommate has made a mess in a

common area. Each scenario also specified parties that were harmed and could potentially sanction the selfish behavior: (a) the child's siblings, the child's parents, and the other parents; (b) the head nurse, and another nurse with a degree from a prestigious school; (c) a roommate, and a roommate's visitor.

For each specified party respondents judged the appropriateness of that party punishing/reprimanding the selfish behavior. Similar judgments were made also of the alternative sanction of praising/rewarding the others who had not acted selfishly. Judgments were made on a five point scale (-2=*highly inappropriate*, -1=*somewhat inappropriate*, 0=*neither appropriate nor inappropriate*, 1=*somewhat appropriate*, 2=*highly appropriate*). It was made clear that the appropriateness of sanctions should be judged in comparison to the party not reacting at all, so a response below zero means that the respondent thought the party in question should *not* sanction.

Finally, respondents were asked for each scenario whether "situations more or less like this scenario (where your answer would be the same) are common," using a five-point response scale from -2=*very uncommon* to 2=*very common.*

Results

The majority of respondents thought that situations similar to those in the scenarios were quite common or very common (the two highest points on the five-point response scale): 66% in the making-a-mess scenario, 64% in the coming-late scenario, and 66% in the eating-the-sweets scenario. An ANOVA for each scenario revealed no significant differences between countries. We conclude that these everyday social dilemmas were recognized across cultures as commonly occurring.

Table 2 summarizes judgments of appropriateness of the fourteen possible combinations of parties and sanctions. The overall result is that most combinations are judged as inappropriate. In other words, most parties should *not* engage in sanctions. However, in every scenario there was one clear exception, i.e., one particular party for whom it was clearly appropriate to use one particular sanction. For instance, in the making-a-mess scenario it was inappropriate for the visitor, but appropriate for the roommate, to punish the messy roommate. Similarly, in the coming-late scenario it was inappropriate for another nurse with a degree from a prestigious school, but appropriate for the head nurse, to punish the late-coming nurse. Finally, in the

eating-the-sweets scenario three different levels arose: it was inappropriate for another parent, weakly appropriate for a sibling, but clearly appropriate for the child's parent, to punish the sweet-eating child. Interestingly, use of *rewards* was typically judged as inappropriate for *all* involved parties in all scenarios (with the exception of the head nurse for which it was weakly appropriate to use rewards).

In summary, the general pattern of judgments across the three scenarios was that rewards were generally inappropriate and punishments were inappropriate unless delivered by a context-specific preferred party: the roommate, the head nurse, and the child's parent, respectively. To quantify this pattern, three domain indices were computed for each participant: *reward by any party* (average judgment of 7 items; α = 0.64; Mean ± SD = -0.27 ± 0.93); *punishment by a non-preferred party* (average judgment of 3 items; α = 0.57; Mean ± SD = -0.68 ± 0.91); *punishment by a preferred party* (average judgment of 3 items; α = 0.81; Mean ± SD = 1.18 ± 0.81). The item "punishment by the child's sibling" was excluded as this item lay somewhere between non-preferred and preferred.

Our main interest lies in punishment by a non-preferred party. An ANOVA with country as a factor showed that the negative mean value (-0.68) was significantly different from zero, $F(1,527)=105.2$, $p<0.0001$, and the effect of country was small and insignificant, $F(72,527)=1.3$, $p=0.06$. Indeed, the country mean was positive just for seven (out of 73) countries and none of these seven countries were represented by more than three participants.

As an additional post hoc analysis, we examined whether individual differences in attitudes to sanctions reflected the three index domains. A principal component analysis of the fourteen judgments was performed. The scree plot supported a three factor solution, explaining 52.5% of the variance. As can be seen in the factor loadings (Table 3), these three factors clearly corresponded to the index domains.[ii] With a factor loadings cutoff value of 0.35, thirteen items each uniquely loaded on the predicted factor. The remaining item was "punishment by the child's sibling", which, consistent with this item's unclear status between non-preferred and preferred, cross-loaded on both punishment-related factors. These results support

[ii] Factor intercorrelations were low (between .03 and .22).

the validity of the index domains and indicate that individuals' sensitivity to roles of punishers and use of rewards generalize beyond specific scenarios.

Discussion

This survey demonstrated that respondents across the world recognized our scenarios, which were chosen to represent everyday social dilemmas. Moreover, in all these everyday social dilemmas and across countries we found evidence of general agreement on informal restrictions of sanctioning rights. We conclude that social norms on use of informal sanctions tend to restrict punishment rights to individuals in certain roles, specific to the context. This is consistent with prior research on legitimacy of social control (Nugier et al., 2007; Chaurand & Brauer, 2008). Of course, our study used only three scenarios. While these three scenarios all yielded the same response patterns, it cannot be ruled out that other broad classes of scenarios exist where norms would be different. Thus, it is still an open question how generally restrictions of sanctioning rights apply.

Our survey also indicated that the content of context-specific restrictions were similar between cultures, such that across different countries the same roles tended to be excluded as punishers. However, this study was limited to a highly selective convenience sample of the world's population (individuals who take part in online studies in English). The tendency of specific restrictions to be universal must be studied further.

Finally, some readers may think it self-evident that it is inappropriate for some parties to punish in the scenarios we used. After all, if there are ingrained norms that not everyone has the right to punish, this is the expected reaction. It is all the more remarkable that the existence of such norms is not reflected in the current literature on punishment and cooperation.

General Discussion

In this paper we have made the following argument: Because of flaring emotions, many people are willing to punish others who behave in ways they dislike – but it is not to the group’s benefit that

punishment is free for all. For one thing, it leads to costly over-punishment. It therefore makes sense for social norms to restrict punishment rights to selected individuals. This is a very different story from the literature that views punishment plainly as a public good that individual needs to be encouraged to contribute to (e.g., Henrich & Boyd, 2001).

We presented two studies investigating different parts of the argument. First, a public goods experiment showed that a condition where each group member's punishment capability was restricted to just one designated punishee yielded higher profits than a condition where every group member could punish everyone else. This extends the scope of positive effects on profits from restrictions of who can punish who, which had previously been investigated for the special case of centralization to a single punisher (O'Gorman et al., 2009). While each study is not very strong evidence on its own, our study and the study of O'Gorman and colleagues together constitute more robust evidence that restrictions of who can punish who reduces over-punishment while still sustaining high levels of cooperation. The additional importance of our finding is that it demonstrates that centralization is not necessary for restriction of punishment rights to increase efficiency. Of course, very many different non-centralized punishment structures are possible. Although a more systematic experimental investigation of different structures would be desirable, we think the logic captured in the model presented in the introduction always apply.

Note that our experiment did not manipulate the legitimacy of designated punishers; every participant was simply designated as a punisher of someone else. Research on legitimacy of authorities has shown that perceived legitimacy, in particular a shared sense of legitimacy, creates voluntary compliance with norms (Zelditch & Walker, 1984; Tyler, 1997). For experiments of the type considered here, this would suggest that efficiency could be even higher if, by some manipulation, punishers were to be perceived as more legitimate.[iii]

[iii] For instance, although centrally restricted punishment was shown to be successful at increasing contributions, both by O'Gorman et al. (2009) and Baldassarri & Grossman (2011), the cost of punishment was still so high that experimental treatments without punishment yielded higher profits. This could be because the abstract and anonymous experimental design does not give punishers in the laboratory the same legitimacy that designated punishers would typically have outside the laboratory.

Our second study was an international survey that presented three scenarios where someone behaved selfishly in everyday social dilemmas. Participants judged for various involved parties whether it would be appropriate or inappropriate for them to punish the selfish behavior. Across many different countries we found that it was judged as inappropriate for most parties to punish. In other words, we found strong evidence for social norms restricting sanctioning rights exclusively to individuals in certain roles.

Our two studies together demonstrate that informal restrictions of punishment rights exist and that they can be beneficial to a group. However, as we mentioned in the introduction we do not think any kind of group selection is necessary to explain the origin or prevalence of these norms. Here we sketch a process that relies only on learning. The core of our argument is that individuals will try to avoid having to suffer obvious downsides of the use punishment. First, even if the intended consequence of punishing someone is to deter selfish behavior for the benefit of the group, the actual consequences may be poor both for the punisher (who may be subject to anger and retaliation) and for the group (which may suffer from poorer group morale, whereas the selfish behavior may not be deterred at all). Second, the consequences are likely to be different depending on who punishes who in what situation. In any given situation where there is heterogeneity in roles and relationships – including power, authority, trust and responsibility in relation to a potential punishee – some potential punishers are more likely to use punishment successfully (i.e., to achieve high compliance and avoid negative effects). Further, the abovementioned kinds of consequences of punishment are easily observable. Individuals may therefore note patterns of successes and failures in use of punishment and form intuitions about who should and should not use punishment in various situations. These intuitions may then feed back into the roles and relationships in terms of further adding to the legitimacy of some roles as punishers and further reducing the legitimacy of others. Because perceived legitimacy in itself is a determinant of compliance, this will reinforce the differences between those roles that tend to be more successful, and therefore preferred, as punishers, and the others who tend to be less successful, and therefore non-preferred, as punishers.

Of course, the selection of a preferred punisher and the rejection of non-preferred punishers rely on some heterogeneity in roles to begin with.[iv] In lab experiments on punishment in social dilemmas, heterogeneity is typically absent by design (participants are anonymous and games are perfectly symmetric). We predict that ideas about who can be sanctioned by whom would develop if heterogeneity were to be introduced, for instance by labeling players as "Parent A", "Child A", "Parent B" and "Child B" and letting participants interact within parent-child dyads before and after the social dilemma part of the experiment. This is for future research to test.

It should be noted that norms that restrict punishment to a context-dependent legitimate punisher may address several of the problems with the efficiency of punishment. For instance, the problem with counter-punishment (Nikiforakis, 2008) is potentially mitigated by norms that proscribe the use of counter-punishment against a legitimate punisher while allowing counter-punishment against an illegitimate punisher, such that a parent may counter-punish someone who punished her child. Similarly, the problem with antisocial punishment (Herrmann et al., 2008) is potentially mitigated by norms that strip antisocial punishers of their legitimacy. Both these possibilities are interesting areas for future investigations into norms restricting punishment.

Finally, note that our survey shed new light on the use of positive sanctions like rewards and praise. Some experimental research has shown that rewards may work better than negative sanctions as a solution to the welfare problem in public goods games (Rand et al., 2009). The results of our survey suggest that this solution to social dilemmas may have limited scope outside the laboratory, because it was generally considered inappropriate to react to selfish behavior by rewarding people who were not selfish. We did not expect this finding; our theoretical argument deals with punishment only and rewards were included in the survey for the purpose of completeness of sanctioning options. According to a recent meta-analysis, rewards and punishments can both be effective to raise cooperation (Balliet, Mulder & Van Lange, 2011). We have

[iv] In real situations where there is no heterogeneity, surveys indicate that it is largely perceived as inappropriate for any individual to punish, whereas collectively organized punishment is judged appropriate (Strimling & Eriksson, forthcoming).

no ready explanation of the finding of inappropriateness of rewards. Why and when it is considered inappropriate to use rewards are important questions for further research.

Acknowledgments

This research was supported by the Swedish Research Council [grants 2009-2390 and 2009-2678].

References

Amster, R. (2003). Restoring (dis)order: Sanctions, resolutions, and "social control" in anarchist communities. *Contemporary Justice Review: Issues in Criminal, Social, and Restorative Justice, 6,* 9–24.

Andreoni, J. (1988). Why free ride? Strategies and learning in public good experiments. *Journal of Public Economics, 37,* 291–304.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review, 80,* 1095-1111.

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Science, 108,* 11023–11027.

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin,* 137, 594-615.

Carpenter, J. P. (2007). Punishing free-riders: how group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior, 60,* 31–51.

Chaurand, N., & Brauer, M. (2008). What determines social control? People's reactions to counternormative behaviors in urban environments. *Journal of Applied Social Psychology, 38,* 1689-1715.

Cinyabuguma, M., Page, T., & Putterman L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics, 9,* 265-279.

de Quervain, D. J.-F., Fischbacher, U., Trever, V., Schellhammer, M., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science, 305,* 1254–1258.

Dreber, A.,Rand, D. G.,,Fudenberg, D. & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*, 348-351.

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

Henrich, J., & and Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208,* 78-89.

Herrmann, B., Thöni, C. & Gächter, S. 2008. Antisocial punishment across societies. *Science, 319,* 1362–1367.

Lerner, J. S., & Tiedens, L. Z. (2006). Portrait of the angry decision maker: how appraisal tendencies shape anger's influence on cognition. *Journal of Behavioral Decision Making, 19,* 115–137.

Netter, B. (2005). Avoiding the shameful backlash: Social repercussions for the increased use of alternative sanctions. *Journal of Criminal Law & Criminology, 96,* 187–215.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: can we really govern ourselves? *Journal of Public Economics, 92,* 91–112.

Nikiforakis, N., & Normann, H.-T. (2008). A comparative statics analysis of punishment in public goods experiments. *Experimental Economics, 11,* 358–369.

Nugier, A., Niedenthal, P. M., Brauer, M., & Chekroun, P. (2007). Moral and angry emotions provoked by informal social control. *Cognition and Emotion, 21,* 1699-1720.

O'Gorman, R., Henrich, J. & Van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B, 276,* 323-329.

Rand, D. G., Dreber A., Ellingsen, E., Fudenberg, D., & Nowak, M. (2009). Positive interactions promote public cooperation. *Science, 325,* 1272-1275.

Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., & Frith, C.D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439,* 466–469.

Strimling, P., & Eriksson, K. (forthcoming). Regulating the regulation: Norms about how people may punish each other. van Lange, P., Yamagishi, T., Rockenbach, B. (eds.), *Social Dilemmas: Punishment and Rewards.*

Tyler, T. R. (1997). The psychology of legitimacy. *Personality and Social Psychology Review. 1,* 323–344.

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

Zelditch, M., Jr., & Walker, H. A. (1984). Legitimacy and the stability of authority. *Advances in Group Processes, 1,* 1-25.

Appendix 1: Instructions to Study 1

You are going to play a game, over several rounds, in a group of four participants. In each round, every participant receives 20 points. From these 20 points you will be able to choose how many points you want to put in a common pot. In the same way, every member of your group will choose how much to put into the pot.

When all participants have decided their contribution to the common pot, the amount in the pot will automatically be doubled. The contents of the pot will then be divided equally to each participant, regardless of their previous contribution. This means that for each point you or anyone else contribute to the pot, you will receive 0.5 points back from the pot. (You will therefore make a net loss of half a point for each point you contribute.)

This game was played for eight rounds. After each round participants were informed of others' contributions. (The others were labeled A through C along the cycle, see Figure 1) . The experiment then proceeded to the punishment treatment, where the previous instructions were repeated with the following addition (boldfaced part depending on whether the condition was restricted or unrestricted punishment):

After each round you will be able to see how much everyone has contributed and you will then have a possibility of punishing **one player/the other players**. To punish others costs you points, but it costs the punished player even more. For each point you spend on punishing someone, that player loses three points. For example, if you spend five points on punishing someone, that player will lose fifteen points. You can spend up to ten points on punishing another player.

This game was played for eight rounds. After each round participants were informed of others' contributions. In the restricted punishment condition, they were then asked how many points they wanted to spend on punishing "player A"; in the unrestricted punishment condition, they were asked the same question for each of players A through C. Finally, they were informed about any punishment they received (but they were not informed about how others were punished).

Appendix 2: Questionnaire used in Study 2

Below we present three scenarios where someone behaves selfishly with respect to a group. For each scenario we are interested in your opinion on:

- *who* can legitimately act in reaction to this behavior, and
- whether appropriate actions are to *punish/reprimand the person who misbehaved* or to *reward/praise those who didn't misbehave* (or both).

Specifically, we will present some possible combinations of who reacts and whether the reaction is by punishing or rewarding. For each such reaction scenario we will ask your opinion on whether it is more or less appropriate than no reaction at all. The following response scale will be used:

- *2: Highly appropriate compared to no reaction at all.*
- *1: Somewhat appropriate compared to no reaction at all.*
- *0: Neither appropiate nor inappropriate compared to no reaction at all.*
- *-1: Somewhat inappropriate compared to no reaction at all.*
- *-2: Highly inappropriate compared to no reaction at all.*

At a gathering of two families, one of the children (say, Kevin) has eaten up the sweets that everyone in the two families were supposed to share after dinner. Both families are around when this is found out.

a) How appropriate would it be if the other parents punish/reprimand Kevin in this situation?

b) How appropriate would it be if Kevin's siblings punish/reprimand Kevin in this situation?

c) How appropriate would it be if Kevin's parents punish/reprimand Kevin in this situation?

d) How appropriate would it be if Kevin's siblings reward/praise everyone but Kevin in this situation?

e) How appropriate would it be if the other parents reward/praise everyone but Kevin in this situation?

f) How appropriate would it be if Kevin's parents reward/praise everyone but Kevin in this situation?

g) Would you say that situations more or less like this scenario (where your answer would be the same) are common?

□*very uncommon* □*quite uncommon* □*neither common nor uncommon* □*quite common* □*very common*

At the hospital, one nurse (say, Rachel) did not show up until very late one day; in the meantime the others had to work extra hard. Both other nurses, of varying educational backgrounds, and the head nurse (Rachel's supervisor) are around when this is found out.

a) How appropriate would it be if another nurse on the ward, with a degree from a more prestigious school than Rachel, punishes/reprimands Rachel in this situation?

b) How appropriate would it be if the head nurse punishes/reprimands Rachel in this situation?

c) How appropriate would it be if another nurse on the ward, with a degree from a more prestigious school than Rachel, rewards/praises everyone but Rachel in this situation?

d) How appropriate would it be if the head nurse rewards/praises everyone but Rachel in this situation?

e) Would you say that similar situations, where your answer would be the same, are common?

□*very uncommon* □*quite uncommon* □*neither common nor uncommon* □*quite common* □*very common*

In a student apartment, one of the students who live there (say, Cath) has created a mess. Both a roommate to Cath and a visitor to the roommate are around when this is found out.

a) How appropriate would it if Cath's roommate's visitor punishes/reprimands Cath in this situation?

b) How appropriate would it be if Cath's roommate punishes/reprimands Cath in this situation?

c) How appropriate would it be if Cath's roommate's visitor rewards/praises everyone but Cath in this situation?

d) How appropriate would it be if Cath's roommate rewards/praises everyone but Cath in this situation?

e) Would you say that similar situations, where your answer would be the same, are common?

□*very uncommon* □*quite uncommon* □*neither common nor uncommon* □*quite common* □*very common*

Figure 1. The implementation of restricted punishment rights in Study 1.

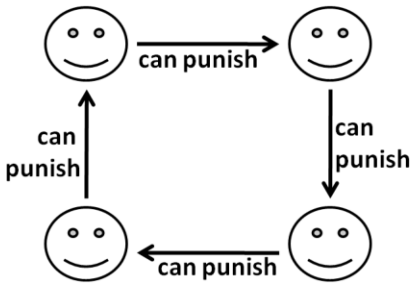

Figure 2. Mean profit per round in each condition.