Regulating the regulation: Norms about punishment

Pontus Strimling and Kimmo Eriksson

Abstract

Rules about punishment dictate how one must behave to ensure that one's punishment behavior is not met with social disapproval. These rules can be both prescriptive, telling us when we have to punish and how much we must punish at a minimum, and restrictive, telling us when we cannot punish or what the maximum punishment can be. In this chapter we investigate the general features of these rules, focusing on punishment of norm violations in social dilemmas. Researchers have often viewed the provision of punishment as a costly public good that must itself be enforced, creating a second order social dilemma that requires prescriptive norms for people to "cooperate", i.e., to punish. We argue that this is a misunderstanding of the nature of punishment and go through theoretical reasons for why prescriptive rules about punishment might not be important. Instead, we discuss the reasons that *restrictive* norms could benefit the group and review experiments where this is shown to be the case. Finally we report the results of four surveys that use real world situations to assess people's views about punishment in several countries. We find that punishment behavior is regulated by generally agreed upon views (i.e., norms), which are largely restrictive rather than prescriptive. Results show a strong consistency across scenarios and countries, indicating that these norms follow general principles.

This is a postprint version of: Pontus Strimling and Kimmo Eriksson (2014). Regulating the regulation: Norms about punishment. In van Lange, P., Yamagishi, T., Rockenbach, B. (eds.), *Reward and Punishment in Social Dilemmas.* Oxford University Press, pp. 52-69.
DOI:10.1093/acprof:oso/9780199300730.003.0004

Introduction

Rules that govern behavior are ubiquitous within every human society and include both formal rules written down in law books and *norms*[1]. Transgressions against these rules can be punished to ensure that the transgressor goes back to acting according to the rules. These punishments are in and of themselves behaviors that are governed by rules of their own and the structure of these punishment related rules is the topic of this chapter.

There are two important questions about norms surrounding punishment: First, if it is costly for the individual to punish, how do groups make sure that punishment is carried out? Second, if punishment is costly for the group, then how do groups limit that cost? These two questions lead us to investigate two different types of rules: prescriptive rules which will tell people what they should do and restrictive rules that tell people what they are not allowed to do.

A special set of situations where punishment is believed to be important are *social dilemmas*. These are situations where individual interests are in conflict with the interest of the group. These situations are important and provide the context for the vast majority of research on sanctions, both punishment and rewards. Although the main focus of this chapter is on social dilemmas, it should be noted that behavioral rules cover a host of different strategic situations and that we see no reason to believe that the structure of rules surrounding punishment would differ depending on the underlying situation. For example, traffic violations are punished whether they are social dilemmas (for instance speeding) or coordination games (which side of the road to drive on).

[1]We define norm as an established standard of behavior shared by members of a social group to which each member is expected to conform

While there are studies of what formal punishments people believe are fair when it comes to people breaking the laws of a society (Hamilton and Rytina 1980, Warr et al 1983), we have found no prior work on norms about informal punishments. This is surprising, not least because there is much research on how norms govern behavior in many situations including social dilemmas (e.g., by Robert Cialdini and his colleagues), and there is some research on the use of informal sanctions (e.g., by Markus Brauer and his colleagues). The aim of this chapter is to present our initial explorations of what norms about informal sanctions look like. Our focus is on the extent to which rules about punishment in social dilemmas are prescriptive or restrictive. Before we present the empirical findings we will review the theoretical arguments for and against prescriptive and restrictive rules.

Reasons for and against prescriptive rules

In the case of social dilemmas there is a well-known theoretical argument that makes a prediction about what norms will surround punishment. Punishment is assumed to be costly to the punisher, at the very least in terms of the opportunity cost of not spending the time and energy on something else. Because of this cost, individually rational agents will not punish voluntarily. Assuming punishment to be directed against agents who do not behave in the interest of the group, punishment will deter selfish behavior and thereby lead to an increase of the group's welfare in the long run. Thus, the provision of punishment is a public good, but because it is costly it will not be provided by agents unless there is a second level of norms regulating that you must punish (or you will be punished yourself). This is known as the second order problem (e.g., Elster 1989; Yamagishi 1986). This argument leads to the prediction that, in the context of social dilemmas, the general character of norms about punishment will be prescriptive, making it obligatory to punish those who do not act in the interest of the group.

There are several reasons to be doubtful about the assumptions behind this prediction. Consider the assumption that punishment will not be provided unless somehow enforced. This assumption has been shown to be invalid in a large number of experiments in which participants in social dilemmas have been given the option to contribute to the punishment of others at a cost to themselves; a robust finding is that a substantial proportion of participants voluntarily do so (e.g., Fehr & Gächter 2002; Ostrom, Walker & Gardner 1992; Yamagishi 1986). Thus, it seems that no enforcement is needed for a substantial proportion of people to be willing to engage in punishment of others.

Now consider the assumption that the provision of punishment is a public good. Data relevant to this issue exist from many laboratory studies of punishment in the public goods game (PGG). Participants in this game receive an endowment in each round. This endowment can be kept as private property or invested in a common good that is equally beneficial to all group members, regardless of their contribution. After each round participants can choose to spend some of their resources on punishing others. The general finding is that contributions go up when punishment is introduced. However, because of the costs of punishment this increase in contributions does not automatically translate to increased group payoff. An extensive meta-analysis of these studies only finds tentative evidence for an increase in group payoff and only several rounds after punishment was introduced (Balliet et al. 2011). Furthermore, since the choice of targets of punishment is left up to the individual, there is no guarantee that it will be directed against those who contributed least to the common good. Experiments in some countries find that punishment is often used against those who contribute much to the common good, which tends to result in contributions no higher than in a no-punishment treatment and a negative effect of punishment on group payoff (Herrmann, Thöni & Gächter 2008). Finally, even when there is an advantage of having

punishment in the first place, this can be eradicated if the punished party has the possibility of taking revenge for being punished (Nikiforakis 2008). To summarize these findings in laboratory experiments, the provision of punishment on a voluntary basis often has bad consequences for the group payoff and there is no general evidence that punishment is a public good in these settings.

Interestingly, the theoretical view of the goodness of voluntary punishers in public goods games is typically not shared by the actual participants of experiments. When people only play one round of PGG they rate punishers negatively on a range of personality features including trustworthiness and likability (Kyonari and Barclay 2008). However, when the subjects play several rounds of PGGs they start to evaluate the punishers as more trustworthy but not as nicer (Barclay 2006). When there was an opportunity to directly reward or punish the players who punished in previous rounds, they were actually given *less rewards* and *more punishment* than the people who did not punish; these results were not always significant but they were consistent across three experiments (Kyonari and Barclay 2008). In an experiment by Cinyabuguma et al. (2006) where each round of PGG was followed by two rounds of punishment, first stage punishers attracted more, not less, second stage punishment. (It should be noted that some of the second stage punishment in this latter study could be due to revenge; however, revenge is not a possible explanation for the results of Kyonari and Barclay.)

Even if punishers are not directly rewarded it is possible that they could receive indirect rewards by being a preferred partner in other games from which they might benefit. Testing this idea, Horita (2010) had subjects read about a PGG in which there was one punisher and one non-punisher and decide which of these individuals they would rather be partnered with in a series of games. Preferences were found to differ between two classes of games. In games

where *the partner would be in a position to provide resources to the participant* (e.g., if the partner was to be the dictator in a dictator game with the participant as the recipient[2]), the punisher tended to be the preferred partner. In contrast, in games where *the participant would be in a position to provide resources to the partner* (e.g., if the partner was to be the recipient in a dictator game with the participant as the dictator), the non-punisher tended to be the preferred partner. Thus, it seems that punishers are not preferred partners in those games where they could be rewarded.

While all of the abovementioned studies centered on punishment in a public goods game, Rob Nelissen (2007) looked at how people viewed third party punishers in a dictator game[3]. He found that people saw these punishers as more fair, friendly, and generous and that they were entrusted with more money in a trust game. Thus, it would seem that the view of third-party punishers is more positive than the view of peer punishers in PGG.

In conclusion, studies of how punishers are viewed indicate that they are generally not well liked and are not rewarded for being punishers. A notable exception is the study of third-party punishers. Our interpretation of this discrepancy is that third-party punishers have been given a special role whose task it is to mete out punishment, whereas in PGG all participants are peers. We shall return to this very important distinction below.

Reasons for and against restrictive rules

If the decision to break the rules is determined by a calculation of the risk of getting caught times the cost of being punished, one might expect that more severe punishment would make people cheat less. However, there are good reasons not to believe in such a simple

[2] In a dictator game one player (the dictator) is endowed with tokens while the second player (the recipient) is not. The dictator can then choose how he wants to distribute those tokens between himself and the recipient.

[3] Third party punishment means that a third player can punish the dictator by removing some of his tokens after the distribution decision has been made.

relationship. One obvious reason is that people often do not behave as rational agents – but even under the assumption of rational agents the total strategic structure may be such that the degree of punishment does not affect equilibrium levels of cheating (Eriksson and Strimling 2012; Weissing and Ostrom 1991). Further, even if there was a monotone relationship between punishment degree and the amount of rule breaking, it might still be that higher punishment is undesirable. One reason is that it is typically desirable that the punished party becomes a productive member of the group, which could be impeded by too severe punishment. Another reason is that the punished party will sometimes be wrongfully accused, in which case all of the cost to the punished is a waste. People may have good reason to doubt that those who voluntarily punish others have made an impartial and knowledgeable assessment of the long-term benefits for the group. After all, people may well make the alternative interpretation that voluntary punishers are angry and act on a personal desire to punish. Such an interpretation would be consistent with psychological research on anger and punishment. For instance, people tend to act on anger whether it is rational or not to do so (Lerner & Tiedens 2006). Brain imaging studies show that people may derive personal satisfaction from punishing others (de Quervain et al. 2003; Singer et al. 2006). We shall return to how people view punishers in one of the studies we report below.

One way of restricting punishment rights is to centralize them to a single agent, regardless of who is to be punished. Laboratory experiments have shown that restriction of punishment rights increases the efficiency of use of punishment, both when rights are centralized to one punishing agent (Baldassarri & Grossman 2011; O'Gorman, Henrich & Van Vugt 2009) and when rights are decentralized so that every agent punishes one other agent (Eriksson, Strimling & Ehn, in press). These findings support the notion that low levels of

punishment are optimal for the group, and that norms that restrict punishment could serve a general purpose of lowering levels of punishment to less destructive levels.

Conclusion

There seems to be no evidence that rules prescriptive of punishment are necessary in order to get some people to punish. Nor is there any evidence that people who choose not to punish are seen as rule breakers. In contrast, restrictions on punishment can increase group benefit. However, the empirical literature is focused on situations in which everyone is allowed to punish. It is possible that if punishment rights are indeed restricted to specific positions, the need for prescriptive rules about punishment increases to make sure that everyone who has the punishment position actually punishes. To investigate this, we below look both at situations where there are specific punishment positions and situations where there are none.

Research from Markus Brauer's group indicates that positions must be taken into consideration in theorizing about punishment: Both potential punishers and potential punishees perceive differences in roles as extremely important in regards to who can legitimately sanction deviant behavior (Chaurand & Brauer 2008; Nugier et al. 2007). Perceived legitimacy, in particular a shared sense of legitimacy, creates voluntary compliance with norms (Tyler 1997; Zelditch & Walker 1984).

As we mentioned previously, there are signs that people’s view of punishers change over the course of several rounds of games played in laboratory experiments. One interpretation of this finding is that people are used to one set of norms and the more time they spend in the laboratory the more time they have to develop new norms. To ensure that

we are studying real-world norms our surveys present situations that are common in the real world rather than abstract games.

Empirical studies of norms about use of informal sanctions

As mentioned, we have found no research focused on norms that regulate the use of informal sanctions in social dilemmas (or in any other situation in which there are norms about behavior). Here we report our own initial explorations in the form of four survey studies. They deal with the following four questions:

1. Is every group member allowed to voluntarily punish a selfish individual or is the right to punish restricted to group members in certain roles?

2. Given that group members in certain roles have the right to voluntarily punish a selfish individual, are they also allowed to choose any level of punishment that they want or are there restrictions on the severity of punishments?

3. When no group member has a special role, how are voluntary punishers regarded compared to those who do not voluntarily punish a selfish individual?

4. When no group member has a special role, does legitimacy of punishment of a selfish individual depend on it being collectively managed rather than individually volunteered?

The studies used 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. Similarly, studies used punishments taken from the realm of punishments available outside the laboratory. To demonstrate that these scenarios tap into the same psychological mechanisms that laboratory experiments do, the third study also included an

abstract social dilemma of the kind used in laboratories, where both payoffs and punishments are given in terms of money.

Surveys were administered online to participants recruited through the Amazon Mechanical Turk (https://www.mturk.com).[4] Users of the Mechanical Turk can come from any country in the world, allowing us to examine norms in more than one culture. Data came primarily from the United States and India, as the vast majority of Turk users are located in these countries. For Study 1 we managed to recruit respondents from many countries across all continents. For Study 4, a Swedish sample of university students complemented the American and Indian Turk users.

Study 1: Are punishment rights restricted to certain individuals?

The first study is taken from Eriksson, Strimling and Ehn (in press); for details, we refer to the original paper. The survey was completed by 528 participants (63% male; mean age 30 years) of mixed educational backgrounds. Based on country of residence, they were divided into six subsamples of unequal size: Asia (N=213), Europe (N=170), North America (N=67), Latin America (N=27), Africa (N=26), Australia and New Zealand (N=25).

The questionnaire presented three scenarios (see Appendix). The first scenario described two families dining together and discovering that one of the children has already eaten the sweets meant for dessert (*depletion of a common resource*); the child could potentially be sanctioned, punished or rewarded, by the child's siblings, by the child's parents, or by the other parents. The second scenario described a hospital ward where one nurse has

[4] Online surveys using the Mturk have been found to be a source of reliable data (Buhrmester, Kwang & Gosling 2011).

come in to work very late which forces the other nurses to work extra hard (*free-riding on a joint effort*); the late coming nurse could potentially be sanctioned by the head nurse or by another nurse with a degree from a prestigious school. The third scenario described a student apartment where one of several roommates has made a mess in a common area (*pollution of a common environment*); the messy roommate could potentially be sanctioned by another roommate or by a visitor.

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.*

For each specified party, respondents judged the appropriateness of that party punishing/reprimanding the selfish behavior, compared to not reacting at all, on a five point scale between -2=*highly inappropriate* and 2=*highly appropriate*. The alternative response of praising/rewarding unselfish behavior was judged on the same scale.

Results of Study 1

For each of the three scenarios, around two thirds of all 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). There were no significant differences between geographical subsamples. In other words, these everyday social dilemmas were recognized across many cultures.

Use of rewards was typically judged as inappropriate for *all* involved parties in all scenarios, with the exception of the head nurse for whom it was weakly appropriate to use rewards. Use of punishment was also judged as inappropriate except for certain preferred parties: the roommate in the making-a-mess scenario, the head nurse in the coming-late

scenario, and the child's parent in the eating-the-sweets scenario (in which the child's sibling was judged somewhere between appropriate and inappropriate as a punisher). To quantify this pattern, three domain indices were computed for each participant: *reward by any party* (average judgment of 7 items; $\alpha = 0.64$; M =-0.27, SD=0.93); *punishment by a non-preferred party* (average judgment of 3 items; $\alpha = 0.57$; M =-0.68, SD=0.91); *punishment by a preferred party* (average judgment of 3 items; $\alpha = 0.81$; M =1.18, SD=0.81). The item "punishment by the child's sibling" was excluded. As illustrated in Figure 1, every geographical subsample judged punishments by non-preferred parties as inappropriate.

FIGURE 1 ABOUT HERE

Study 2: Is the severity of punishment restricted?

The second survey was completed by 100 participants (56% male; mean age 30 years) of mixed educational backgrounds, from United States (N=50) and India (N=50). Respondents were presented with the same three scenarios as in Study 1: making-a-mess, coming-late, and eating-the-sweets (see Appendix). They were told to imagine that the roommate in the making-a-mess scenario, the head nurse in the coming-late scenario, and the child's parent in the eating-the-sweets scenario (i.e., the roles that were identified as preferred punishers in Study 1) reacted to the selfish individual ("S") in either of five different ways:

1. not reacting at all;
2. explaining to S that what S did was wrong;
3. same as (2) but also yelling at S;
4. same as (3) but also slapping S;
5. same as (4) but also beating S with a stick.

For each reaction, respondents were asked how this would affect their view of the punisher: *negative* (coded -1), *neutral* (coded 0), or *positive* (coded +1).

Results of Study 2

Results were similar for all three scenarios. The only reaction that tended to affect people's view of the punisher positively was to just explain that the selfish behavior was wrong. Yelling tended to be neutral in its effect on people's view of the punisher. People's view of the punisher tended to be negatively affected if the punisher did not react at all, and the tendency was even more negative if the punisher used the more severe punishments of slapping or beating. The response pattern is illustrated in Figure 2, showing the mean ratings for each reaction across the three scenarios. Note how responses were very similar between Indian and American participants.

FIGURE 2 ABOUT HERE

Study 3: How are voluntary punishers regarded compared to non-punishers?

The third survey was completed by 200 participants (54% male; mean age 31 years) of mixed educational backgrounds, from United States (N=100) and India (N=100). It focused on situations where group members are not distinguished by any differences in roles, which is the typical setup in laboratory experiments.

The three scenarios from the previous studies were adapted so that there were no cues to distinguish between group members (see Appendix). For every scenario we then described one group member who decided to let the selfish behavior go (i.e., a non-punisher), and another group member who decided to yell about the selfish behavior (i.e., a voluntary punisher). The decision to use yelling as punishment in these scenarios was based on the

finding in Study 2 that yelling tended to be neutrally viewed when done by a preferred punisher.

A fourth scenario, in which a typical abstract social dilemma was described, mimicked laboratory experiments (see Appendix). Each of three participants could give away money that would then be doubled and split between the other two participants, after which decisions they could punish each other by paying an amount to deduct three times the same amount from the punishee of their choice. As in the previous scenarios, the scenario described one selfish person, one non-punisher and one punisher.

The order of scenarios was counterbalanced, so that the experiment scenario was presented either first or last with 100 participants in each condition. For each scenario respondents compared the punisher against the non-punisher on seven traits, on a five-point scale coded from -2 = *definitely [the non-punisher]* to 2 = *definitely [the punisher]*. The seven items were:

1. you would prefer to spend time with;
2. most likely to punish people unfairly;
3. most likely to adhere to standard norms of behavior;
4. most likely to be an angry person;
5. most likely to take others' interests into account;
6. most likely to create bad morale in the group;
7. most trustworthy.

Results of Study 3

There were no order effects, so we present the results for the pooled data. Results were similar for all four scenarios and both countries, as illustrated in Figure 3. Respondents tended to prefer to spend time with non-punishers; they also tended to find non-punishers most likely to adhere to standard norms of behavior, to take others' interests into account, and to be trustworthy. Punishers, on the other hand, tended to be viewed as most likely to punish other people unfairly, to be angry people, and to create bad morale in the group.

FIGURE 3 ABOUT HERE

Study 4: Collectively managed vs. individually volunteered punishment

The fourth survey was completed on paper by 18 Swedish students of computer science and online by 100 participants (54% male; mean age 31 years) of mixed educational backgrounds from United States (N=50) and India (N=50). The aim of this study was to investigate the legitimacy of individual voluntary punishment compared to collectively managed punishment.

The survey presented four variations of a scenario where a group has a joint task that requires multiple meetings. One group member tends to come late to these meetings, and in the end this group member has to buy coffee to everyone in the group. This involves three steps: *decision on the norm* (that it is unacceptable to come late), *decision on the punishment* (that latecomers must buy coffee for everyone in the group), and *execution of the punishment* (ensuring that the latecomer buys coffee for everyone). One variation of the scenario had all these steps managed collectively by the group; the other three variations had a single individual, Eric, voluntarily stepping in instead of the collective, either in the last step or

earlier in the first or second steps (see Appendix). The order of scenario variations was counterbalanced.

Respondents were asked, for each scenario variation, whether they found Eric's behavior to be OK. A response scale from -3 = *definitely not OK* to 3 = *definitely OK* was used. The same question was then asked about the group's behavior.

Results of Study 4

Eight respondents were excluded because they gave the exact same response to all questions, indicating that they had not paid attention. All three countries showed the same pattern: When the individual, Eric, manages every step involved in punishment, his behavior tended to be viewed as not OK (i.e., below the midpoint of zero). Eric's behavior tended to be viewed as more OK when more steps were managed collectively rather than by Eric, see Figure 4. The group's behavior was also viewed as more OK the more steps were managed collectively.

FIGURE 4 ABOUT HERE

Discussion

Across scenarios and cultures, we found remarkable consistency of norms regarding informal punishment. In the two studies with distinguishable roles we found evidence both for prescriptive norms (i.e., views that a certain party ought to punish) and restrictive norms (i.e., views that a certain party ought not to punish) present whereas in the situations without distinguishable norms we found only restrictive norms.

Norms in social dilemmas with distinguishable roles

Studies 1 and 2 investigated social dilemmas in which involved parties had different roles with respect to the person who behaved selfishly. To have the right to punish in these situations, it was not sufficient to be part of the group that suffered from the selfish behavior. Involved parties tended to be normatively constrained from punishing unless they had a special role. In this special role, people would tend to view you negatively if you do not punish at all, but even more so if you use a punishment that is too severe. Therefore, there are both prescriptive and restrictive norms occurring. The socially acceptable behavior is to punish only if you are in the punishing role and for that punishment to be deemed by the group as an appropriate amount, not too little nor too much.

While humans seem to have a knack for picking up norms, there is reason to believe that it is still difficult to learn whether or not you have the right position to punish, or what the right amount of punishment. We expect norms to be inferred from experience to no little extent. Direct experience can be relied on when you infer a norm about a common behavior. However, norm-breaking is typically rare (or else we would not call it a norm) so punishment for norm violation will be rare. In addition, the behavior expected of you is dependent on your position within the situation so you can only rely on the punishment behavior you have observed for that specific position. Thus, for any given situation in which you could potentially punish someone, people will typically suffer a lack of previous experience on how to behave in that particular situation. They can therefore be expected to draw on their experiences from other punishment situations and assume that analogous norms hold also in the present situation. In other words, in order for there to be unspoken norms about punishment these norms need to be easily generalizable or people will simply not know them.

This could explain why there is so little difference in attitude towards punishment between our different scenarios.

Laws about punishment

A special case in which there are distinguishable roles is formal law. There are prescriptive and restrictive principles in criminal laws both now and historically. The idea that every criminal act has both a minimum and maximum punishment is found in all law books, even those of the very first laws in which every law is a description of a specific infringement and the exact punishments it merits (Jarrick and Bondesson 2011).

The law often restricts people from using certain behaviors (e.g., violence) to punish each other. This principle is explicit already in the Tang Code (624 AD), in which it is stated that a person who acts outside of the law to revenge the death of his parents should be punished by lifetime banishment (Johnsson 1997). Conversely, failing to fulfill the punishment duty of one's position carries penalties in most societies that employ an institutionalized justice system.

Norms in social dilemmas with no distinguishable roles

Studies 3 and 4 investigated social dilemmas in homogeneous groups. These surveys showed that individual members who voluntarily punish others are viewed negatively in various ways. Study 4 showed that the preferred alternative is for the group to manage punishment as a collective. This is consistent with anthropological studies of real world punishment in small scale societies (Guala 2012).

While collectively managed punishment may have advantages over individually managed punishment in terms of optimizing the level of punishment, it is clear from Study 3 that efficiency is not the only concern people have about voluntary punishers. Individual

punishers tended to be regarded as angry persons who were not acting in the interests of others and likely to use punishment unfairly and create bad morale in the group. Voluntary punishers were not even regarded as trustworthy in this survey. This is consistent with previous findings of punishers being judged as trustworthy only if participants had played the public goods game several times (Barclay 2006).

The results of our surveys were remarkably constant not just across scenarios but also across cultures. This stands in stark contrast with other norms, such as norms regarding punctuality, where there are often substantial differences between countries (Levine et al. 1980). Of course, it is possible that this universality across cultures hold only for the particular scenarios we happened to use. Even so, the indication that there might be human universals in norms regulating punishment is enough to warrant future studies.

Concluding remarks

In this chapter we have investigated what norms surround punishment. In survey based research we found, throughout the world, the existence of norms that heavily restrict punishment. To a lesser extent we found norms that make punishment obligatory. Situations were viewed differently depending on whether everyone involved were in the same position (colleagues, siblings or friends) or whether someone had an elevated position (a teacher, a parent or a boss). Norms that restrict punishment were found in both cases, but norms that make punishment obligatory were found only in situations where someone had a special position (and then only that person was obligated to punish). The finding of no obligations to punish in the case where there are no positions might to some extent depend on the limited number of scenarios we used. Perhaps there exist other scenarios without special positions where some punishment is seen as obligatory. However, even if this is the case those norms

would have to be different from the norms about situations where people with special positions are expected to hand out punishment.

This leaves would be punisher in a precarious position. They must be sensitive to norms that tell them not to punish too much as well as to norms that tell them not to punish too little, and sensitive to how this depends on their position in the group. The opportunity to see and learn norms surrounding punishment only arise when someone has broken a norm for how to behave in the first place. This gives anyone who is learning punishment norms fewer occasions to learn what the acceptable behavior is compared to other norms. Nonetheless we found remarkable agreement on the norms, not just within countries but also across countries. We suggested that the solution to this paradox might be that punishment norms become generalizable between situations, as people have no choice but to generalize.

The findings in this chapter stand in stark contrast to the notion that punishment in itself is seen as a public good. The most positive interpretation of how people view punishment is as a double edge sword that may benefit the community by ensuring that people adhere to norms but harmful when overused or used by the wrong person. The surveys conducted here do not address whether or not punishers are altruistic or even see themselves as altruistic. Instead they join an increasing number of studies that find that others often view punishers in a negative light. However, in a new experiment we have found that the individuals who punish behavior that is harmful to the group are to a large extent the same individuals who punish behavior that benefits the group (details available from the authors). This suggests a general lack of altruistic motivations among punishers that would help explain *why* people tend to view them so negatively.

References

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

Balliet, D., Mulder, L.B. & Van Lange, P.A.M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin.* 137, 4.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior.* 27, 5, 325—344.

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6, 3-5.

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.

Elster, J. **(**1989). *The cement of society: A study of social order*. Cambridge: Cambridge University Press.

Eriksson, K. and Strimling,P. (2012) The hard problem of cooperation. *Plos ONE*, 7(7): e40325. doi:10.1371/journal.pone.0040325

Eriksson, K., Strimling, P., & Ehn, M. (in press). Ubiquity and efficiency of restrictions on informal punishment rights. To appear in *Journal of Evolutionary Psychology*.

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

Guala, F. (2012). Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and brain sciences,* 35, 1–59.

Hamilton. V. L. & Rytina. S, (1980). Social Consensus on Norms of Justice: Should the Punishment Fit the Crime? *American Journal of Sociology,* 85, 5, 1117-1144.

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

Horita, Y. (2010). Punishers May Be Chosen as Providers But Not as Recipients. *Letters on Evolutionary Behavioral Science.* 1, 6—9.

Jarrick, A. & Wallenberg, M. (2011). Flexible Comparativeness: Towards Better Methods for the Cultural Historical Study of Laws - And Other Aspects of Human Culture, *Organizing History*. Nordic Academic Press,

Johnson, W. 1997, The T´ang Code. Volume II, Specific Articles, Princeton University Press

Kiyonari, T. and Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of personality and social psychology,* 95, 4.

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.

Levine, R.V., West, L.J., & Reis, H.T. (1980). Perceptions of time and punctuality in the United States and Brazil. *Journal of Personality and Social Psychology,* 38, 4.

Nelissen, R. (2008). The price you pay: cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, 29, 242—248.

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

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.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review*, 86, 404-417.

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

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

Warr, M., Meier, R. & Erickson, M. Norms, (1983). Theories of Punishment, and Publicly Preferred Penalties for Crimes. *Sociological Quarterly.* 24, 75—91.

Weissing, F. and Ostrom, E. (1991). Irrigation institutions and the games irrigators play: Rule enforcement without guards. *Game Equilibrium Models II: Methods, Morals, and Markets.* 188—262.

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.

Figure 1. How respondents from different geographical regions judged the appropriateness of sanctions in Study 1. Bars show average judgments, across all three scenarios, of rewards by any party, punishments by a non-preferred party, and punishments by a preferred party.

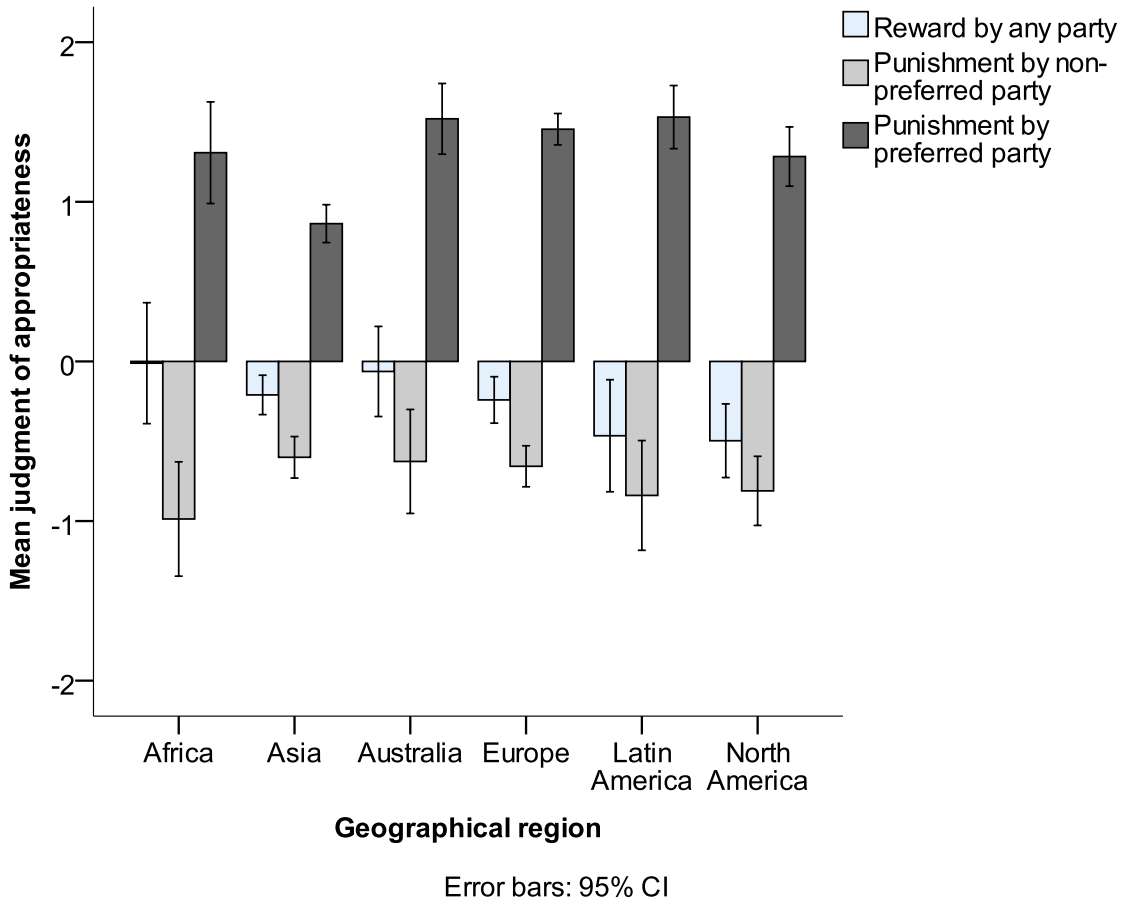


Figure 2. How respondents in Study 2 judged voluntary punishers depending on the severity of the punishment.

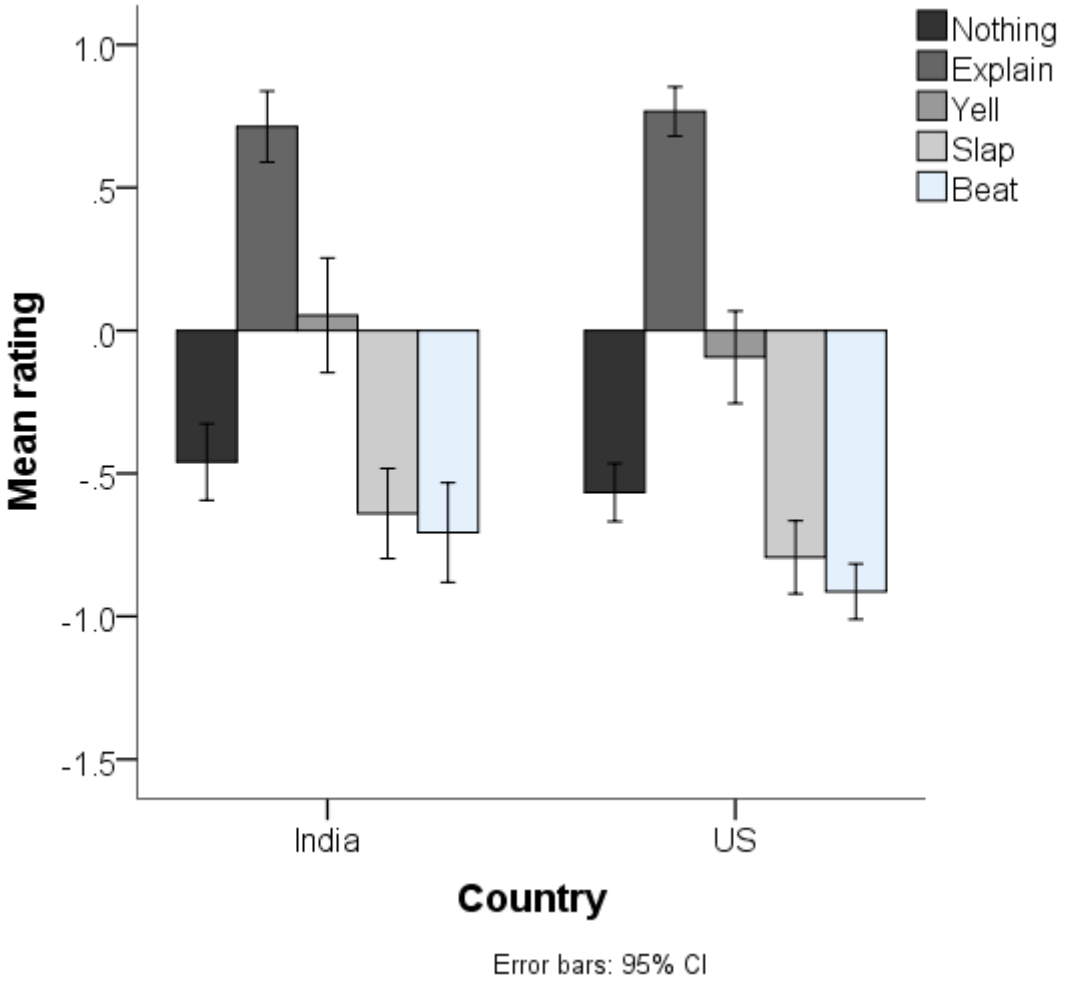


Figure 3. How respondents in Study 3 compared voluntary punishers to non-punishers on seven traits in each of four scenarios. Positive values on an item represent tendencies to view the punisher higher than the non-punisher on the trait.

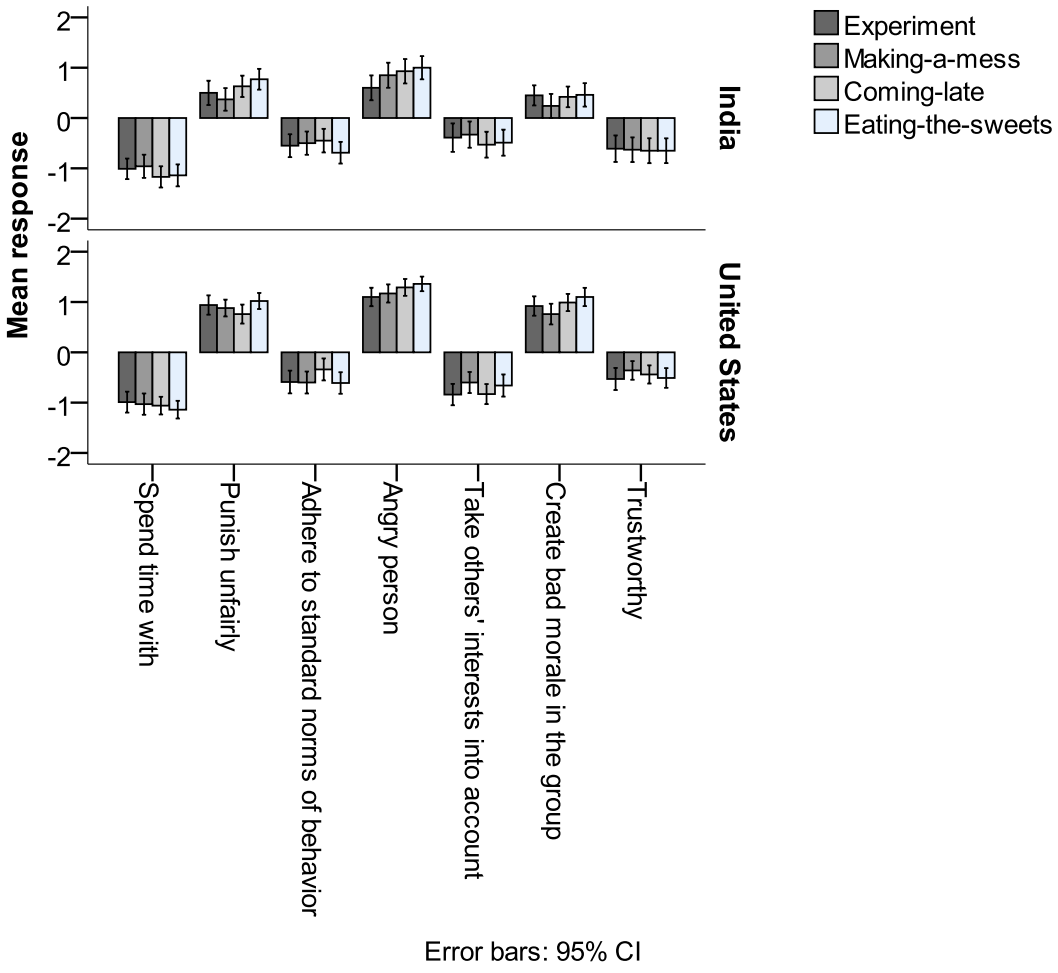
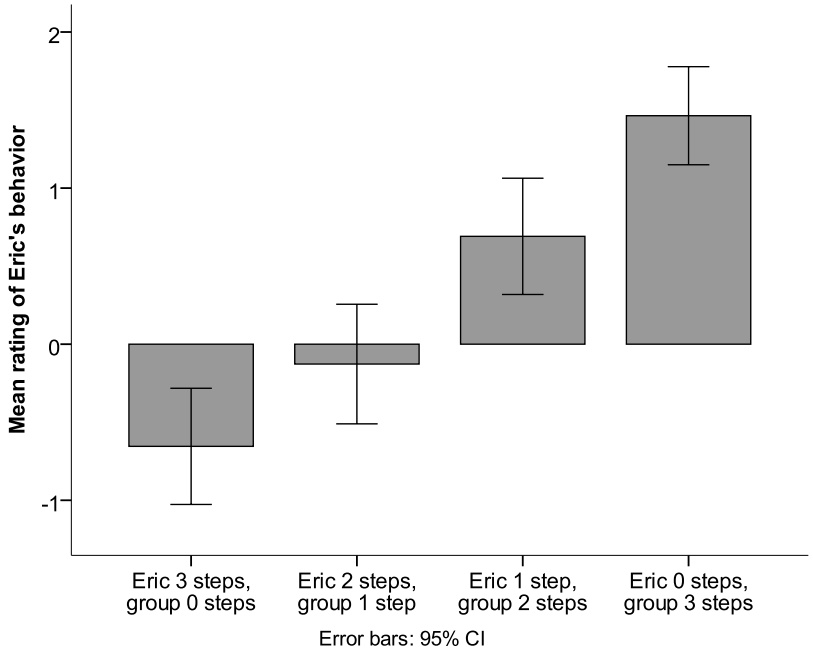
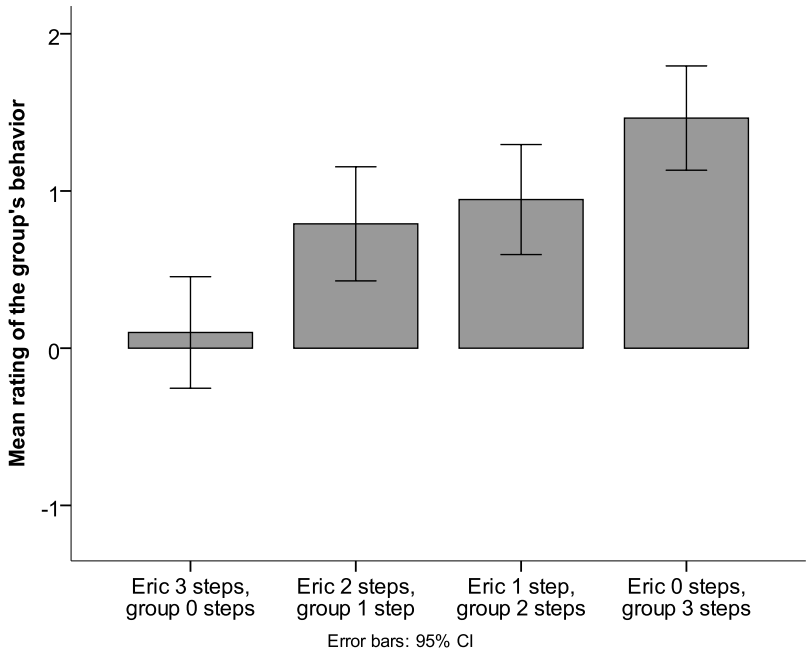


Figure 4. Study 4 judgments of the behavior of the single voluntary punisher Eric (top) and the behavior of the group (bottom) depending on how many of the steps involved in punishment were managed by Eric instead of the group as a collective.





Appendix

Scenarios used in Studies 1 and 2

Eating-the-sweets: At a gathering of two families, one of the children (Kevin) has eaten up the sweets that everyone in the two families was supposed to share after dinner. Both families are around when this is discovered.

Coming-late: At the hospital, one nurse (Rachel) did not show up until very late one day; in the meantime the others had to work extra hard. Other nurses, of varying educational background, as well as the head nurse (Rachel's supervisor) are around when this is discovered.

Making-a-mess: In a student apartment, one of the students who lives there (Cath) has created a mess. Both Cath's roommate and a visitor of the roommate are around when this is discovered.

Scenarios used in Study 3

Eating-the-sweets: At a gathering of a few classmates after lectures, one of them (Kevin) has eaten the sweets that everyone was supposed to share. Upon noticing this, the other classmates have different reactions: Paul decides to let it go whereas Ron decides to yell at Kevin.

Coming-late: At the hospital, one nurse (Rachel) did not show up until very late one day; in the meantime the other nurses in her team had to work extra hard. When Rachel eventually arrived, the other nurses had different reactions: Sarah decides to let it go whereas Maria decides to yell at Rachel.

Making-a-mess: In a student apartment, one of the students who lives there (Cath) has created a mess. Her two roommates have different reactions: Jennie decides to let it go whereas Frances decides to yell at Cath.

Experiment: An economics experiment involves three participants. Everyone is given 10 dollars that they can choose to keep or voluntarily give away to the others. Every dollar they give away is matched by the experimenter. This means that if a participant gives away a certain amount, both the others receive that amount in full. After these decisions have been made, they can sacrifice some money to

deduct from someone else's earnings: For every dollar they sacrifice, three dollars are deducted from the participant of their choice. In the experiment, Carl and Peter both gave away some money, Mark did not give away anything. Carl decides to let it go whereas Peter decides to sacrifice 2 dollars in order to deduct 6 dollars from Mark.

Scenarios used in Study 4

Individual decides norm and punishment and executes punishment: At the first meeting, the group member Eric thinks about the importance of arriving on time and decides for himself that coming late is unacceptable. Eric then finds John and tells him that he has come up with a suitable punishment: Each time John comes late in the future he must buy coffee for the entire group. As it happens, John comes late to a couple of the following meetings, and each time Eric makes sure John buys coffee for the entire group.

Group decides norm, individual decides punishment and executes punishment: At the first meeting, the entire group discusses the importance of arriving on time and jointly decides that coming late is unacceptable. One group member, Eric, then finds John and tells him that he has come up with a suitable punishment: Each time John comes late in the future he must buy coffee for the entire group. As it happens, John comes late to a couple of the following meetings, and each time Eric makes sure John buys coffee for the entire group.

Group decides norm and punishment, individual executes punishment: At the first meeting, the entire group discusses the importance of arriving on time and jointly decides that coming late is unacceptable. The group jointly decides on a suitable punishment for latecomers: Each time John comes late in the future he must buy coffee for the entire group. As it happens, John comes late to a couple of the following meetings, and each time one of the other group members (called Eric) makes sure John buys coffee for the entire group.

Group decides norm and punishment and executes punishment: At the first meeting, the entire group discusses the importance of arriving on time and jointly decides that coming late is unacceptable. The

group jointly decides on a suitable punishment for latecomers: Each time John comes late in the future he must buy coffee for the entire group. As it happens, John comes late to a couple of the following meetings, and each time the other group members (one of whom is called Eric) together make sure John buys coffee for the entire group.