Informal punishment of non-cooperators

Kimmo Eriksson

23 October 2017

Abstract

According to an influential theory known as "strong reciprocity", humans cooperate at high levels due to the rise of *altruistic punishers*, that is, individuals who not only cooperate themselves but also informally punish non-cooperators. Strong reciprocity theory assumes that this punishment is costly to the punisher but beneficial to the group, that is, the punisher behaves altruistically. The theory further assumes that by engaging in this individually costly but group-beneficial behavior, punishers gain a good reputation. The aim of my dissertation is to critically examine the empirical validity of these assumptions through a series of experimental studies. Overall, I find that the assumptions of strong reciprocity theory are not supported. (1) Punishment of non-cooperators does not seem to be driven by punishers having the group's interest at heart. In fact, I find that punishers in economic cooperative games tend not to be more cooperative than non-punishers. Punishers also tend to punish both non-cooperators *and* cooperators. I conclude that punishers seem to be characterized by being generally punitive rather than being generally altruistic. (2) Punishers of non-cooperators do not seem to gain a good reputation in general. Rather, informal social norms about the use of punishment seem to restrict it more than encourage it. Moreover, people who face the choice of whether to punish a non-cooperator seem not to tend to think of punishing as the moral thing to do.

My conclusion of these empirical results is that strong reciprocity theory paints an incorrect picture of the psychology of informal punishment of non-cooperators. I argue that this theory likely goes wrong already when it takes cooperative situations as its starting point, and that a better approach would be to assume that there is a more general psychology of informal punishment. I sketch what such an approach would entail.

Contents

Preface

This is dissertation for a PhD in social psychology based on published works. I am originally a pure mathematician, trained at the Royal Institute of Technology in Sweden. I obtained a PhD in 1993 with a dissertation on algebraic combinatorics. I then extended my research interests to include stable matchings, a topic in the intersection between combinatorics and game theory. In 1999, I was appointed as a dean at Mälardalen University in Sweden. When I went back to research in 2003 I had somewhat lost touch with what was happening in algebraic combinatorics. I decided to focus on game theory instead. Specifically I was interested in the accuracy of game theoretic predictions of matching behavior (see Eriksson & Strimling, 2009).

I soon found out that experimental work on behavior in games was already being conducted by behavioral economists across the world, although there was no dedicated laboratory for such experiments in Sweden. In order to create such a laboratory I joined forces with an economist, Martin Dufwenberg, and a sociologist, Peter Hedström. Our grant application was approved but both my co-applicants were offered jobs at prestigious institutions abroad. I was left in charge of the grant despite having no experience in behavioral experiments whatsoever. To manage this challenge I asked my new PhD student in mathematics, Pontus Strimling, to join me in finding out how to set up a lab and how to conduct experiments. We turned out to be a good team. Pontus has been a valuable co-author on all the research I report in this dissertation as well as many articles not reported here.

Before Peter Hedström moved to Oxford he organized a series of seminars in Stockholm that I attended. One talk was given by the experimental economist Simon Gächter, who presented joint work with Ernst Fehr on the public goods game. This is a game in which each member of a group decides how much money to contribute to a common pot. The total value of the common pot then grows by some multiplicative factor and is then distributed equally to all group members. High contributions are better for the group, but selfish individuals are tempted to make a low contribution. From earlier research on the public goods game, it was well-known that contributions to the pot tended to decline when the game was repeated several times. The novelty of Fehr and Gächter's experiment was that once everyone's contributions were known, players were given the option to pay a small cost to reduce the payoff of any other player. This change in the game made contribution levels rise dramatically.

I remember being impressed by these results—but not by the researchers' sweeping interpretation of them. First, they interpreted the public goods game as embodying the general concept of cooperation; specifically, cooperation was reduced to an individual behavior, namely, to make a large contribution to the common pot. Second, they interpreted their experimental results as showing that the reason humans cooperate better than other related species is that they punish those who do not cooperate. The behavior to punish low-contributors, termed *altruistic punishment*, was interpreted as a biologically determined strategy that had evolved because it enabled humans to cooperate better; in other words, the payoffs in the public goods game were interpreted as corresponding to fitness.

My skeptical reaction was based on reflection on my own experience. For instance, as an amateur musician I enjoyed playing chamber music. Making music together seemed to me a prime example of human cooperation: the players make a joint effort in the creation of a good and the result depends crucially on every player doing his/her part. No good chamber music is made if one musician plays at

another tempo than the others, or in another key, or not at all. Nonetheless, ensemble playing did not seem to fit the above framework. Amateur musicians choose to cooperate (i.e., do ensemble playing as well as they can) because it is its own reward, not because they are afraid of punishment from the other players. Moreover, if another player would yell at someone who plays the wrong note I would certainly not regard that as an altruistic behavior. And I really did not see how the rewards and costs associated with good ensemble playing would affect the players' fitness in an evolutionary sense. For that matter, I didn't see how public goods game payoffs could reasonably be interpreted as fitness either.

Over a number of years I conducted research on a variety of topics related to games and cultural evolution. My original doubts about the notion of altruistic punishers remained, but I did not address them directly in my research. A turning point came in 2011 when I went to the International Conference on Social Dilemmas and met the social psychologist Toshio Yamagishi. He demonstrated that skepticism of the notion of altruistic punishers could be the basis of interesting research. Since then I have conducted a number of studies in this direction. This dissertation is based on five papers of mine that, in various ways, question the notion of punishers as altruistic.

Acknowledgments

The research in this dissertation was funded by the Swedish Research Council and carried out within my employments as a professor at Mälardalen University and a guest professor at Stockholm University. I am grateful to the School of Psychology at the University of Kent for admitting me to the Ph.D. program in social psychology and for assigning me Tim Hopthrow and Robbie Sutton as encouraging supervisors. I tip my hat to my wife Frances for letting me spend time and money on obtaining a second Ph.D. Finally, I thank my co-authors on the five papers for working with me on this research. The contributions of co-authors are summarized below.

Paper 1. Authors: Kimmo Eriksson, Daniel Cownden, Micael Ehn, and Pontus Strimling. KE came up with the original idea for this paper and developed the idea further together with PS. KE designed both studies included in the paper: a lab experiment and a reanalysis of previously published data by Herrmann et al (2008). ME programmed the lab experiment after KE's design and conducted it. KE performed all statistical analyses. KE drafted the paper, but the original draft was significantly improved by DC; in particular, DC wrote the discussions of evolutionary hypotheses. PS gave constructive feedback on the drafted paper.

Paper 2. Authors: Kimmo Eriksson, Pontus Strimling, and Micael Ehn. KE developed the original idea for the paper together with PS. KE designed both studies included in the paper: a lab experiment and a survey experiment. KE conducted the survey experiment. ME programmed the lab experiment after KE's design and conducted it. KE performed all statistical analyses. KE drafted the paper. PS gave constructive feedback on the drafted paper.

Paper 3. Authors: Pontus Strimling and Kimmo Eriksson. PS and KE developed the idea for this paper together. PS and KE also developed the designs of the three new survey studies together. KE conducted the surveys and performed all statistical analyses. KE drafted the paper, but the original draft was substantially and significantly improved by PS; in particular, PS wrote the literature review and the final discussion.

Paper 4. Authors: Kimmo Eriksson, Per A. Andersson, and Pontus Strimling. KE came up with the original idea to use Heider & Simmel-style animations to study judgments of peer punishment. KE developed this idea further together with PS. PA produced the actual animations according to specifications. KE designed all studies. KE conducted all three online studies. PA conducted the lab study. KE performed all statistical analyses. KE drafted the paper. PS and PA gave constructive feedback on the drafted paper.

Paper 5. Authors: Kimmo Eriksson, Pontus Strimling, Per A. Andersson, and Torun Lindholm. KE came up with the idea when working with TL on another project involving the ultimatum game. KE developed this idea further together with PS. KE designed all studies in this paper. KE conducted all studies with the exception of the paid experiment, which was programmed and conducted by PA. KE performed all statistical analyses. KE drafted the paper. PS, PA and TL gave constructive feedback on the drafted paper.

1 Review of the literature that has motivated my research

The notion of altruistic punishment stems from the game theoretic approach to understanding human behavior. Game theory is a branch of mathematical modelling in which situations are represented by a set of players, each facing a choice between several strategies. The payoff to each agent is assumed to be determined by the combination of all their choices.

1.1 Nash equilibria and social preferences

Classic game theory makes the additional assumptions that (1) players make their choices to maximize their own payoff and (2) by some process they will arrive at a combination of choices such that no player can increase their payoff by unilaterally changing strategy. The strategy combinations referred to in the second assumption are called Nash equilibria. Thus, given a social situation where people care about what others do, classic game theory says that a researcher should model it by a game, that is, identify the strategies available to agents and identify the payoffs of every strategy combination. By calculating the Nash equilibria of that game, the researcher could then make predictions about how people will behave in the situation.

In economic experiments in which people play games with monetary payoffs, observed behavior often does not amount to a Nash equilibrium. To try and salvage the Nash equilibrium as a predictive tool, behavioral economists tend to relax the assumption that all players are *selfish,* that is, that their aim is to maximize their own payoff. Some players are instead thought to be *prosocial*, that is, care also about that others get sufficient payoff.

1.2 Social dilemmas and the game theoretic approach to cooperation

A *social dilemma* is a game in which individual players face a choice between promoting their self-interest (usually referred to as 'defection') and promoting the group-interest (usually referred to as 'cooperation'). In other words, in a social dilemma individuals benefit from others' cooperation but not from their own. The working assumption of many game theory oriented researchers of cooperation is that social dilemmas are good models of the value of, and hurdles for, successful real-life cooperation. Specifically, this model allows researchers to think of the value of cooperation as a sum of payoffs, while the hurdle for successful cooperation is simply the selfishness of agents.

I want to emphasize that there are other ways to conceive of cooperation. For instance, the late British social psychologist Michael Argyle explicitly rejected the notion of cooperation as a game theoretic strategy. Instead he defined cooperation as "acting together in a coordinated way at work, leisure or in social relationships, in the pursuit of shared goals, the enjoyment of the joint activity, or simply furthering relationship" (Argyle, 1991). Moreover, lay-people have their own ideas about the hurdles for cooperation. In an unpublished on-line study I asked 100 US participants a checkbox question on which characteristics best describe someone who is difficult to cooperate with. Although many respondents characterized such a person as *selfish* there were several other responses that were equally frequent, including *arrogant, ill-tempered, impatient, inflexible,* and *narrow-minded*. These data suggest that real-life cooperation do not only require that people temper their selfishness, but also that they are able to communicate well and adapt their goals to circumstances.

Other problems with the game theoretic approach to cooperation will be discussed in section 3.1.

1.2.1 The public goods game with punishment

A great variety of different social dilemma games can be obtained from varying the number of players, the number of strategies they have access to, the payoffs, and whether moves are made simultaneously or sequentially. For the purpose of this dissertation it is sufficient to introduce the two games that play the most important roles in the literature on altruistic punishment. One of these is the *public goods game* (PG), which was mentioned in the preface. The PG has multiple (typically four) players. All players make simultaneous decisions on how to share their endowments between self and a common pot. The pot is then doubled in value by the experimenter (or multiplied by some other factor greater than 1 but less than the number of players). After this value increase the pot is distributed equally to all group members. This means that by contributing to the pot all players decrease their own payoff but increase the sum of all players' payoffs. Thus selfish players should contribute nothing whereas sufficiently prosocial players should contribute maximally.

What makes the PG a social dilemma is that individual players can control their own contribution only. From the point of view of payoffs there is no tension between self-interest and group-interest with respect to others' contributions. If players instead had the power to make decisions for another player, both prosocial and selfish players should contribute maximally. Paper 1 in this dissertation describes such an experiment that yielded contributions at high but not maximal level—another indication of the limited predictive power of game theory.

As I mentioned in the preface, Fehr and Gächter (2000) extended the PG with a second stage in which players may select other players (based on their contributions in the first stage) as the target of a payoff reduction. In order to reduce another's payoff, players must accept that their own payoff is reduced too, although by a smaller amount (typically a third). Fehr and Gächter termed the use of such payoff reduction as "altruistic punishment". The logic behind the term "punishment" is that it is an act that has negative consequences for the target of the act. The logic behind the term "altruistic" is that although the punishment only has negative effects on payoffs in the short term, it may deter future defection and thereby maximize the group's future payoff; as such, it could be regarded as future-oriented altruism, if altruism is defined as *benefitting the group at a direct cost to oneself*. Note that this is equivalent to the definition of the cooperative strategy in a social dilemma. (I think it is unfortunate that the terms altruism and cooperation are used as synonymous in this literature, and that neither usage is aligned with how these terms are used in other contexts.)

1.2.2 The ultimatum game: another example of costly punishment

The *ultimatum game* (UG) is a two-player game in two sequential moves. The first player, the *proposer,* decides on an offer of how to share an endowment with the other player, the *responder*. The responder then decides whether to accept or reject the offer. In case of rejection, the endowment is taken back by the experimenter. A selfish responder should therefore always accept any share greater than zero. Rejection of offers that are less than half but better than nothing has been interpreted as altruistic punishment (Fehr & Fischbacher, 2003). The logic behind this interpretation is that offers of

less than half can be regarded as unfair, and rejection of unfair offers could be regarded as punishment of the proposer (as the proposer loses the endowment), which could deter future unfair behavior. Note that fairness in distributive decisions does not benefit the group in the sense that the total amount of resources increases. Nonetheless, the total utility derived from these resources may increase. To see this, consider food-sharing. The benefit to an individual of having more food than he or she needs is low compared to the benefit an individual who needs the food would gain from it. Thus, a fair redistribution of food may increase the benefits derived from the food.

1.3 The strong reciprocity hypothesis of the evolution of human cooperation

Papers on altruistic punishment in economic games have had great impact on the literature on cooperation. The reason is that economic games have been tied into a grand theory that incorporates both evolution and social norms. This theory is usually referred to as strong reciprocity (Fehr, Fischbacher & Gächter, 2002). It has been promoted in a series of papers and books by a collective of researchers, prominent members of which include Ernst Fehr, Herbert Gintis, Samuel Bowles, Joe Henrich, and Robert Boyd. Strong reciprocity theory takes for granted that cooperation can be modeled as a social dilemma, that the costs and benefits of cooperation translate into biological fitness, and that the choice of behaving cooperatively or not is genetically determined. Under these assumptions, cooperation with strangers should not evolve because it would be outcompeted by selfish strategies. But, it is claimed, humans do cooperate with strangers at much higher levels than comparable animals. Hence there is an "evolutionary puzzle of human cooperation". The solution offered is *strong reciprocity*, which means that people do not only cooperate with strangers, they also punish strangers for not cooperating. The strategy that combines cooperating with punishing non-cooperators is referred to as an *altruistic punisher* strategy. The existence of the altruistic punisher strategy in a population changes the game such that selfish strategies are less successful. However, this solution creates a second-order question: Why isn't the altruistic punisher strategy outcompeted by a non-punisher strategy? According to strong reciprocity theory, the solution is the same here: as altruistic punishment is a form of cooperation, non-punishers are punished too (e.g., Boyd, Gintis, Bowles & Richerson, 2003; Henrich & Boyd, 2001).

1.4 The most important experimental studies of altruistic punishers

There is a vast experimental literature on the ultimatum game and the public goods game. Here I discuss just a few papers that I consider to be the most important and relevant ones. A much richer treatment can be found in a rather recent volume on punishment in social dilemmas edited by Van Lange, Rockenbach, and Yamagishi (2014).

1.4.1 Studies of targets of punishment in public goods game experiments

In experiments on the public goods game the players choose whom to target with punishment. Thus there is no guarantee that it will be directed against those who contributed least to the common pot. Indeed, a study conducted in many countries found that in some countries punishment often targeted high contributors too (Herrmann, Thöni & Gächter, 2008). Consistent with a deterrent effect of

punishment, such “anti-social punishment” tended to lead to lower contributions. Thus, punishment can have both positive and negative effects on the group outcome.

Another study around the same time examined the PG with punishment when it was extended by a second round in which the punished party could use counter-punishment against the punisher (Nikiforakis, 2008). The presence of counter-punishment eradicated the positive effect of punishment. Thus, positive effects of punishment seems to be contingent not only on which country you are experimenting in but also on a very specific design of the game.

A related study used a second round of punishment in which any player (i.e., not just punished parties) could direct punishment at other players based on how they used punishment in the first round (Cinyabuguma, Page & Putterman, 2006). Even then, first round punishers received more, not less, second round punishment than first round non-punishers. This is an indication that strong reciprocity theory may be wrong about why people use costly punishment: It does not seem like using punishment makes you less likely to be punished by others.

This conclusion was reinforced by a study that used both punishment and rewards as well as personality ratings (Kiyonari & Barclay, 2008). Punishers received more punishment and less rewards than non-punishers. Moreover, punishers were rated as less likable and less trustworthy than non-punishers.

Note that all these studies have been conducted in the context of economic games in laboratories. Ultimately the altruistic punisher hypothesis concerns human behavior in everyday life. There is work in social psychology studying informal punishment of norm violations in everyday life (e.g., Chaurand & Brauer, 2008). However, the question of how use of punishment is judged by observers has not been studied in that line of work. The work I present in Papers 2–4 focuses on this question.

1.4.2 Studies examining whether ultimatum game rejection is altruistic punishment

The conception of ultimatum game rejection as altruistic punishment has been examined in two studies by Yamagishi and colleagues. In one study the researchers modified the UG such that responders could only reject their own share but not affect the proposer’s share (Yamagishi et al., 2009). Despite not being able to cause any harm to the proposer, rejections of unfair splits were still common in the experiment. This suggests that rejection of an unfair offer may be psychologically distinct from punishment. The researchers suggested that rejection is instead related to asserting one’s independence.

In a later paper the same research group studied correlates of using rejection in the ultimatum game and obtained three important findings (Yamagishi et al., 2012). First, use of rejection was positively correlated with a measure of assertiveness, consistent with the motive suggested in the first study. Second, use of rejection in the UG was uncorrelated with a measure of participants’ vengefulness, again suggesting that a desire for punishment was not the main driver of rejections. Third, use of rejection was uncorrelated with various helping behaviors in the domain of economic games, suggesting that rejection is not driven by an altruistic disposition. These findings are related to the questions I address in Papers 1 and 5.

2 Summary of my research: questions, methods, results, and conclusions

According to a recent review of strong reciprocity theory there is quite compelling evidence that norms of fairness can underlie punishment among humans but "it remains contentious whether this is an altruistic concern for the values of one's group" (Wenzel & Okimoto, 2016, p. 241). The research in this dissertation consists of five papers that were all motivated by my doubts of the notion of "altruistic punishment" and in various ways they address this point of contention. Three big questions are addressed in this research:

1. Those who punish selfish behavior are called altruistic punishers – but are they altruistic? (Paper 1.)
2. What do social norms about punishment of selfish behavior look like? (Papers 2, 3, 4.)
3. Do people think using altruistic punishment is the moral thing to do to? (Paper 5.)

Each of these big questions can be broken down into several more specific questions. Below I describe the specific questions I have addressed in my papers, the methods that were used, and the results that were obtained.

2.1 Those who punish selfish behavior are called altruistic punishers – but are they altruistic?

As outlined in section 1.4, the altruistic punisher in the strong reciprocity theory is a strategy in a social dilemma that combines cooperating with punishing non-cooperators. The empirical basis of this notion is that some participants in the public goods game pay a cost to punish non-cooperators. But to qualify as altruistic punishers, these participants should also be cooperators themselves and they should limit their punishment to non-cooperators. Do they? In Paper 1 (Eriksson, Cownden, Ehn & Strimling, 2014) we addressed these questions by reanalyzing the dataset from Herrmann et al.'s (2008) 16-country study of 1,120 participants who played the public goods game in groups of four.

Strong reciprocity theory in its most basic form views the altruistic punisher simply as a strategy and does not specify what makes an individual follow this strategy. One possibility is that altruistic punishers are simply motivated by the collective interest, both when they cooperate and when they punish. However, in the public goods game it is both in the collective interest and in a player's selfish interest that others cooperate. Thus, another possibility is that punishment is motivated by players' selfish interest. To tease these motives apart we tweaked the game such that only the collective interest would be a motive to use punishment, whereas the selfish interest would motivate restraint on punishment. We conducted an experiment on this tweaked game, also reported in Paper 1.

2.1.1 Do punishers of non-cooperators cooperate themselves?

Method. We reanalyzed the data on the public goods game with and without punishment from Herrmann et al. (2008). In the condition without punishment we computed each participant's total contribution; this is our measure of cooperation. In the condition with punishment we computed the relative frequency with which each participant used opportunities to punish low contributors (those who made below-median contributions); this is our measure of use of 'altruistic punishment'. Within each group of four participants playing the public goods game we could then calculate the correlation

between cooperation and use of 'altruistic punishment'. If punishers of non-cooperators are cooperators themselves, these correlations should tend to be positive.

Results. Contrary to the notion of 'altruistic punishers' in strong reciprocity theory, the mean within-group correlation between cooperation and use of 'altruistic punishment' was zero. A histogram (Fig. 1, left panel) clearly shows that the distribution of within-group correlations was symmetric around zero, illustrating the lack of a tendency for punishers of non-cooperators to be cooperators themselves.

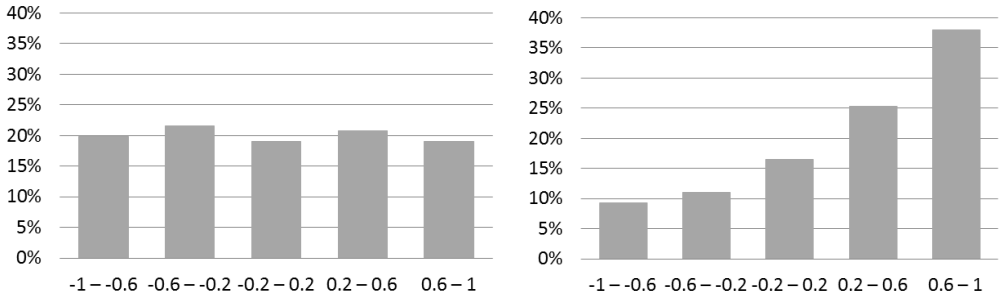


Figure 1. (Reproduced from Eriksson, Cownden, Ehn & Strimling, 2014). Distributions of the within-group correlations between individual use of punishment against low contributors and (left) individual contributions in condition without punishment, (right) use of punishment against high contributors.

2.1.2 Do punishers of non-cooperators avoid punishing cooperators?

Method. For every participant in Herrmann et al. (2008) we computed the relative frequencies with which the participant used opportunities to punish low contributors ('altruistic punishment') and opportunities to punish high contributors ('antisocial punishment'). Within each group of four participants we could then compute the correlation between these two measures. According to the conception of altruistic punishers in strong reciprocity theory, frequent users of 'altruistic punishment' should not be frequent users of 'antisocial punishment'.

Results. Contrary to the notion of 'altruistic punishers' in strong reciprocity theory, the mean within-group correlation between use of 'altruistic punishment' and use of 'antisocial punishment' was positive. The histogram in the right panel of Fig. 1 shows how the distribution of correlations was clearly skewed towards positive values.

2.1.3 Do punishers of non-cooperators use punishment in a way that promotes others' cooperative behavior even when it is not in their selfish interest to do so?

Method. We conducted an economic experiment in which the standard public goods game, with and without punishment, was supplemented by the addition of conditions in which contributions were vicarious: each group member decided the contribution of the next one (cyclically). When contributions are vicarious, it is in each player's selfish interest to deter the vicarious donor to make high contributions on their behalf – whereas it is in the collective interest to promote high contributions. Among those who in the standard game only ever used punishment against *low* contributors (i.e., seemingly pure 'altruistic punishers') we computed whether they used punishment against high or low contributions made vicariously on their behalf. If punishment of non-cooperators

is motivated by desire to promote the collective interest, punishers should not tend to change their targets of punishment from low to high contributors only because it is no longer in their selfish interest to punish low contributions.

Results. Contrary to the notion of pure 'altruistic punishers' being motivated by the collective interest, most of them changed their targets of punishment when the setting changed to vicarious contributions. In other words, the majority punished their vicarious donor for making a *high* contribution on their behalf, thereby deterring cooperation in this game.

2.1.4 Conclusion

The results from these investigations point to a straightforward conclusion: participants who use 'altruistic punishment' in public goods game experiments do not exhibit the other altruistic behaviors in the same domain that would be predicted by strong reciprocity theory. As reviewed earlier some research on altruistic punishment has used the ultimatum game instead of the public goods game, conceiving of rejection of low offers as a form of altruistic punishment. Also rejection of low offers in the ultimatum game is uncorrelated with various helping behaviors in economic games (Yamagishi et al., 2012). In sum it seems that results from experiments on economic games – the very kind of data that initially motivated the altruistic punisher hypothesis – do not support the notion that those who punish selfish behavior have altruistic motives.

Of particular interest is the consistency with which participants used punishment across distinct situations: The same individuals tended to be the most frequent punishers of both high and low contributors, and of both voluntary and vicarious contributions. This consistency is remarkable given that the downstream effects of punishment on payoffs, to self and to others, tend to be positive in some situations and negative in others. Thus there seems to be an individual propensity to use or not use punishment, which is not particularly attuned to its downstream effects on payoff.

In relation to everyday experiences this is not very surprising. We all know some people who seem to be generally more prone to angry confrontation than some other people we know. Nonetheless, it is bad news for the game theoretic approach to understand human behavior. The fundamental assumption of game theory is that behavioral choices can be predicted from people's preferences over the likely consequences of their choices. If people's behavior is shaped not so much by their thoughts about possible consequences as by unrelated emotions and habits, the fundamental assumption of game theory is invalid. It would be particularly problematic for evolutionary game theoretic models of cooperation. Such models rely on the assumption that the evolution of cooperative behavior can be studied within one specific strategic situation. If general proximate mechanisms, such as anger, influence punishment behavior across various situations, then the payoffs associated with any given situation cannot be used to model the evolution of punishment.

2.2 What do social norms about punishment of non-cooperative behavior look like?

Because of the group-beneficial consequences that punishment of selfish behavior may have, many scholars have regarded the provision of punishment as a public good in itself (e.g., Henrich & Boyd, 2001; Nakao & Machery, 2012; Yamagishi, 1986). Accordingly, the strong reciprocity theory count

non-punishers as non-cooperators that will be punished by the altruistic punishers, whereas punishers come into good standing in the group (Henrich & Boyd, 2001). Against this theoretical argument stand the experimental studies of public goods games that found that punishers were *not* socially rewarded relative to non-punishers (Cinyabuguma et al., 2006; Kiyonari & Barclay, 2008). This may not be so surprising, given our abovementioned findings of punishers not exhibiting any clear altruism. If other people question punishers' altruism, it should make them less inclined to reward punishers.

This tension in the literature led me to ask what social norms actually surround informal punishment of non-cooperators. In contrast to the lack of nuances in a typical game theoretic model, actual social norms must deal with the many subtleties in the social world. To begin with, individuals may have many different relations to each other: in the same group there may be both relatives and non-relatives, both superiors and underlings, both roommates and visitors, etc. A first question is whether the social norm is that anyone should punish someone who behaves selfishly, or whether the informal duty to do so is restricted to group members that have a special role in relation to the person who behaves selfishly (like a parent, a superior, or a roommate).

The social world is also different to simple game models in that possible punishments range from extremely mild to extremely severe. Assuming a person in a certain role has the informal duty to punish someone else for behaving selfishly, do social norms restrict how mild or severe the punisher may be?

Not all social structures involve special roles and relations. Some real-life situations are analogous to the public goods game experiments in that they involve a group of *peers*, that is, a group in which there are no relevant distinctions among its members. When there are no relevant distinctions between group members to guide who should punish someone who behaves selfishly, do social norms simply mandate *any* group member to be a punisher? Or is everyone discouraged from punishing? Or should the group members act as a collective?

Moreover, the answers to these questions may depend on various factors. For instance, whereas economic laboratory experiments tend to focus only on monetary punishments, the social world offers a variety of *kinds* of punishment (e.g., yelling, fines, and beatings). Do norms about how peers may use punishment differ between different kinds?

Individual variation should also be considered. As social norms are not explicitly spelled out and agreed upon, we should expect substantial variation between individuals in their social judgments of peer punishers. In particular, assuming that use of peer punishment is often motivated by anger rather than altruism (as we discussed at the end of the previous section), a more positive view of peer punishers could be found among people who are themselves high on trait aggression.

2.2.1 When someone behaves selfishly, is the informal duty to punish restricted to people in certain special roles in relation to the target of punishment?

Method. In Paper 2 (Eriksson, Strimling & Ehn, 2013, Study 2) we conducted an online survey of 528 respondents across the world, recruited through the Amazon Mechanical Turk (mturk.com). The questionnaire presented three everyday scenarios involving someone behaving selfishly in a group whose members had different relations to the selfish person. One scenario described two families dining together and discovering that one of the children has already eaten the sweets meant for

dessert. Another scenario described a hospital ward where one nurse is very late coming in to work, forcing the other nurses of various ranks to work extra hard. A third scenario described a student apartment where someone has made a mess in a common area to the annoyance of both a roommate and a visitor. (These situations were selected as everyday versions of three basic social dilemmas: depletion of a common resource, free-riding on a joint effort, and pollution of a common environment.) For each scenario, respondents judged how appropriate it would be for each group member to "punish/reprimand" the selfish person.

Results. The relation between the selfish person and the punisher had a large effect on the appropriateness of using punishment. Specifically, it was generally viewed that the greedy child should be punished by its parent but not by a grownup in the other family; the late-coming nurse should be punished by the head nurse but not by a peer; the messy student should be punished by a roommate but not by a visitor. The same pattern held in every geographical subsample, as illustrated in Fig. 2.

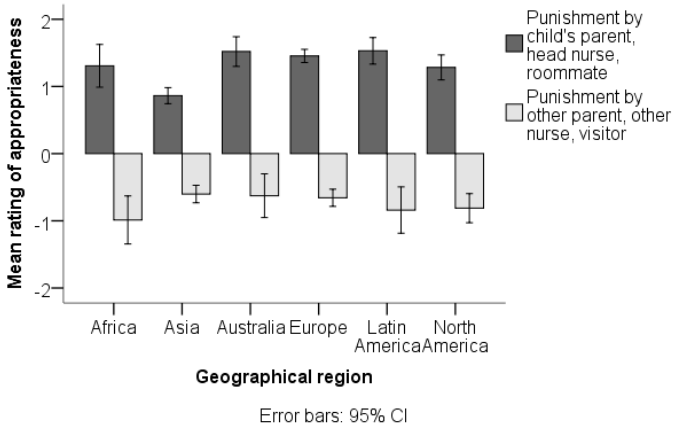


Figure 2. (Based on data from Strimling & Eriksson, 2014). How respondents from different geographical regions rated the appropriateness of different group members punishing/reprimanding someone who behaved selfishly. Ratings were averaged across the three scenarios described in section 2.2.1.

2.2.2 Do social norms restrict how mild and how severe the punisher may be?

Method. In Paper 3 (Strimling & Eriksson, 2014, Study 2) we conducted an online survey of 100 respondents, 50 each from the United States and India. Respondents were presented with the same three scenarios as in the previous study and told to imagine that the child's parent, the head nurse, and the roommate, either did not reacted at all or punished the selfish person in the scenario in either of four different ways: (1) by explaining that the selfish behavior was wrong, (2) by explaining and yelling, (3) by explaining, yelling, and slapping, or (4) by explaining, yelling, slapping, and beating with a stick. For each reaction, respondents were asked how this would affect their view of the punisher: negatively (coded -1), neutrally (coded 0), or positively (coded +1).

Results. Across scenarios the view of the punisher tended to be negatively affected if the punisher did not react at all, indicating the people in these special roles in these situations have an informal duty to react with some kind of punishment. However, a positive effect on people's view of the punisher was obtained only when the punisher just explained that the selfish behavior was wrong. Yelling had no systematic effect, whereas the addition of physical punishments had a clear negative effect on people's view of the punisher. Thus it seems that social norms require a punisher to use as

mild a punishment as possible. Fig. 3 illustrates how these results held both among US and Indian participants.

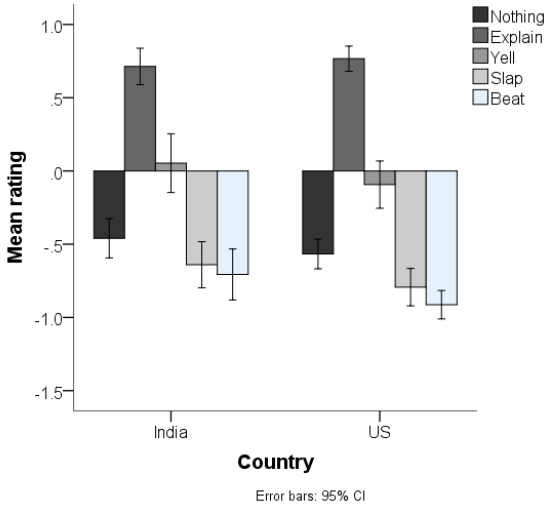


Figure 3. (Reproduced from Strimling & Eriksson, 2014). How respondents from the United States and India rated how use of different punishments would affect their view of a punisher with an informal duty to punish someone who behaved selfishly. Ratings were averaged across the three scenarios described in section 2.2.2.

2.2.3 Within a peer group, is it appropriate for a peer to punish another peer?

Method. In Paper 3 (Strimling & Eriksson, 2014, Study 3) we conducted an online survey of 200 respondents, 100 each from the United States and India. The three scenarios from the previous studies were adapted so that there were no relevant distinctions between group members, and such that every scenario described one peer did not punish and another peer who yelled about the selfish behavior. The decision to use yelling as punishment was based on the finding in the previous study that yelling tended to be judged neutrally when done by a preferred punisher in a group with non-peers. A fourth scenario instead presented a public goods game with one player using monetary punishment. For each scenario respondents compared the punisher against the non-punisher on seven items, on a five-point scale coded from -2 = *definitely [the non-punisher]* to 2 = *definitely [the punisher]*. The seven comparisons 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; and (7) most trustworthy.

Results. Results were remarkably consistent. The non-punisher was judged more favorably than the punisher on all items, across all four scenarios, and in both countries. As illustrated in Fig. 4, the two items showing the largest effect were that the punisher was most likely to be angry person whereas people would prefer to spend time with the non-punisher.

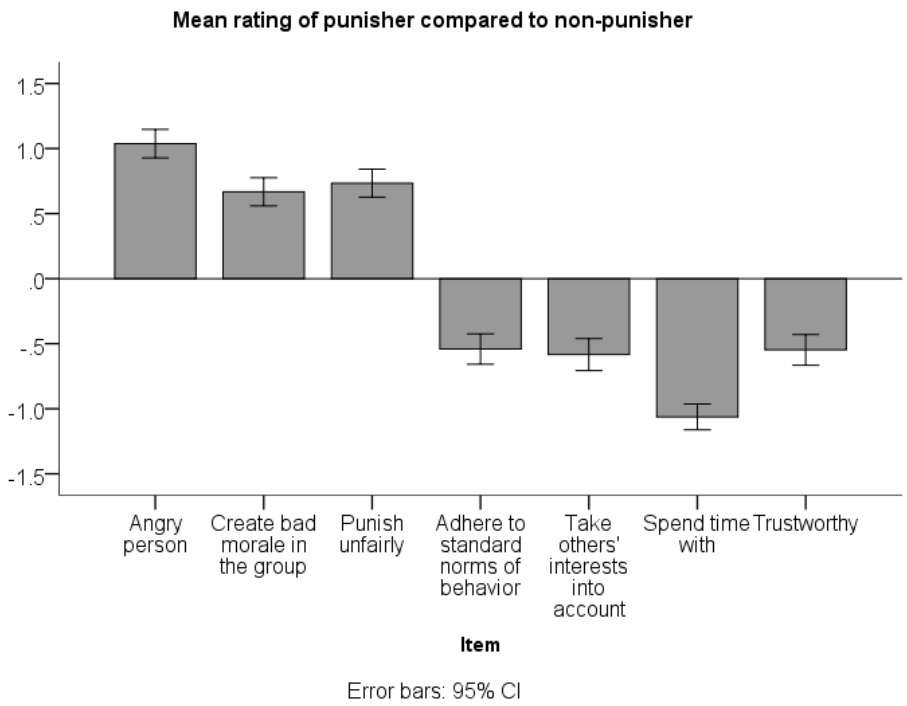


Figure 4. (Based on data from Strimling & Eriksson, 2014). Mean ratings of a punisher compared to a non-punisher. Ratings were averaged across the four scenarios described in section 2.2.3.

2.2.4 Is it more appropriate for a peer group to punish as a collective?

Method. In Paper 3 (Strimling & Eriksson, 2014, Study 4) we conducted a survey of 118 respondents from Sweden, the United States and India. A basic scenario was presented in which a group has a joint task that requires multiple meetings and one group member keeps coming late. The latecomer was punished by having to buy coffee to everyone in the group. This involved 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). In four variations of the scenario the first steps (between zero and three) were managed collectively by the group and the remaining steps were managed by a single group member called Eric. For each of the four variations of the basic scenario, respondents were asked to rate the appropriateness of Eric's behavior as well as the groups behavior (on a response scale from -3 = *definitely not OK* to 3 = *definitely OK*).

Results. All three countries showed the same pattern of results: When all punishment steps were managed collectively, both Eric and the group were judged to act appropriately. When the same punishment was managed singlehandedly by Eric, he was judged to act inappropriately. Intermediate variations yielded intermediate results, as illustrated in Fig. 5.

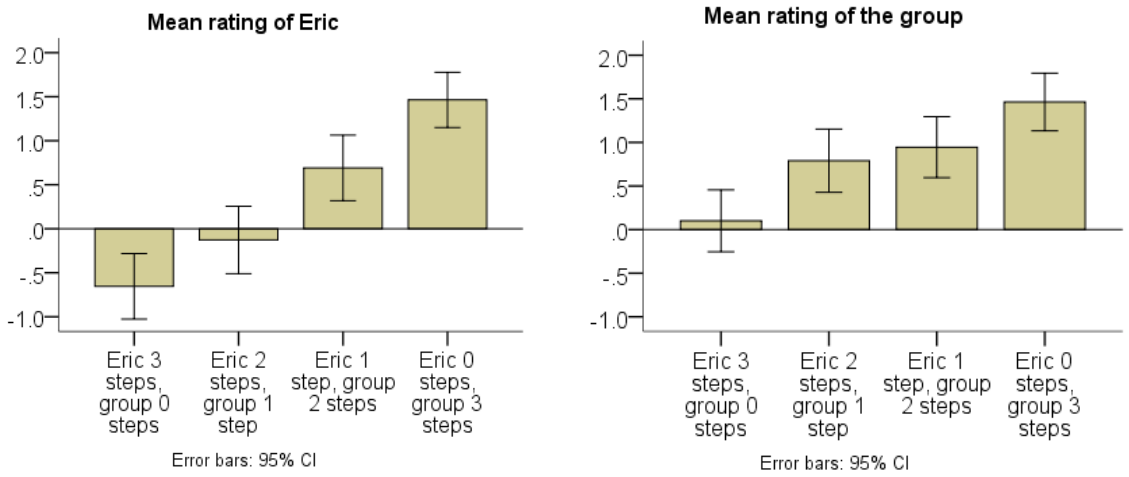
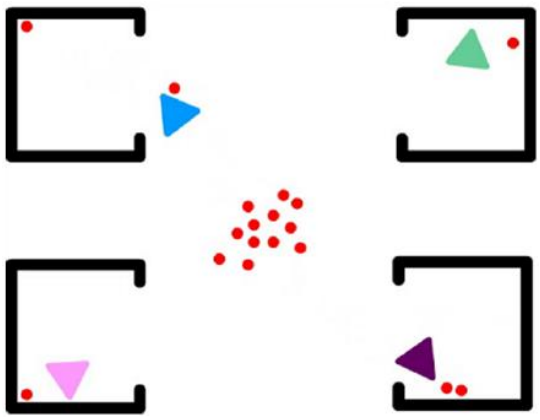


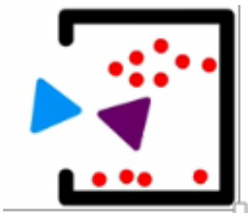
Figure 5. (Based on data from Strimling & Eriksson, 2014). Mean appropriateness ratings of the individual punisher Eric (left) and the group (right) depending on how many steps of the punishment were managed by the individual instead of the group.

2.2.5 *Are social norms about punishment of selfish behavior similar across different kinds of punishment?*

Method. In the previous studies we used written scenarios to study social norms about punishment of selfish behavior. In Paper 4 (Eriksson, Andersson & Strimling, 2016) we developed a method that uses animations instead. This method was inspired by the classic psychological research of Heider and Simmel (1944). Using animations of triangles and circles, these researchers demonstrated that movements are sufficient to make viewers perceive abstract geometric shapes as having intentions, emotions, and social interaction. When I was exposed to this animation in a research seminar, it struck me how it could be advantageously used to study social judgment. Compared to written scenarios, geometric seem to offer several potential benefits. First, watching a scenario being played out rather than reading about it may make it more salient. Second, geometric shapes may be less susceptible to idiosyncratic associations than descriptions of human characters. Third, the weak reliance on language should make geometric animations suitable for cross-cultural and developmental studies. Compared to films of real human actors, geometric animations are also cheap and easy to make using readily available software. The animations can therefore be created in many variants to enable examination of the impact of various situational factors upon social judgments.

In Paper 4 we created a basic animation of four triangles in different colors: blue, green, pink, and purple. The animation shows the triangles engaged in harvesting a resource, represented by a pile of circles. The triangles are strictly observing a sharing norm by taking turns at harvesting one circle at a time. Suddenly the purple triangle violates this norm by harvesting the entire remaining resource in one go. We created a large number of different continuations to this basic animation. One variant shows the other





triangles noting that the resource has been harvested in its entirety but not doing anything about it. In all other variants the blue triangle seeks out the purple triangle and engages in some kind of punishment. The main aim of the study was to examine how social judgments of this peer punisher varied with the details of the punishment, and whether there was any way punishment could be executed to make the peer punisher judged more positively than a non-punisher.

We created animations that depicted various kinds of punishment: *physical aggression* (e.g., a shove), *economic punishment* (taking back resources from the purple triangle), and *yelling* (represented by an exclamation mark in a speech bubble).

In a conceptual replication of the findings from Paper 3, the animations of physical aggression and economic punishment each came in four versions that varied the *severity* of the punishment (a non-harming shove vs. a damaging blow, or taking back some vs. taking back all of resources), and independently varied whether the punishment was *individually or collectively managed* (i.e., whether the blue triangle acted alone or in concert with the green and pink triangles).

Each animation was judged by at least 100 American participants recruited through Amazon Mechanical Turk. The physical aggression animations were also judged by a total of 162 Swedish university students. Each participant watched only one version of the animation and rated the triangles on a number of items. The focus was on the blue triangle, which was rated on eight items: (1) I think the BLUE triangle's behavior was appropriate. (2) I would like to spend time with a person who behaves like the BLUE triangle. (3) [*reverse-coded*] If a person who behaves like the BLUE triangle belonged to my group I would consider that person to be a problem, rather than an asset, for the group. (4) I think BLUE is someone who (a) is generally trustworthy, (b) [*reverse-coded*] is generally angry, (c) takes others' interests into account, (d) [reverse-coded] would punish others unfairly, (e) generally follows standard norms of behavior. Responses were z-score transformed and averaged to a punisher approval index of high internal consistency. Participants also completed a trait aggression scale (Bryant & Smith, 2001).

Results. As illustrated in Fig. 6, the study of US participants replicated the findings of Paper 3: First, peer punishment consistently gained worse approval ratings than non-punishment. Second, individually managed punishment gained worse approval ratings than when the same punishment was collectively managed. Third, more severe punishment gained worse approval ratings than milder punishment. Importantly, these patterns held across physical aggression and economic punishment. Moreover, individual differences were found: the approval of non-punishment and the disapproval of peer punishment were particularly apparent among participants who scored low on trait aggression. See Fig. 7. The same pattern of results were obtained in the Swedish sample.

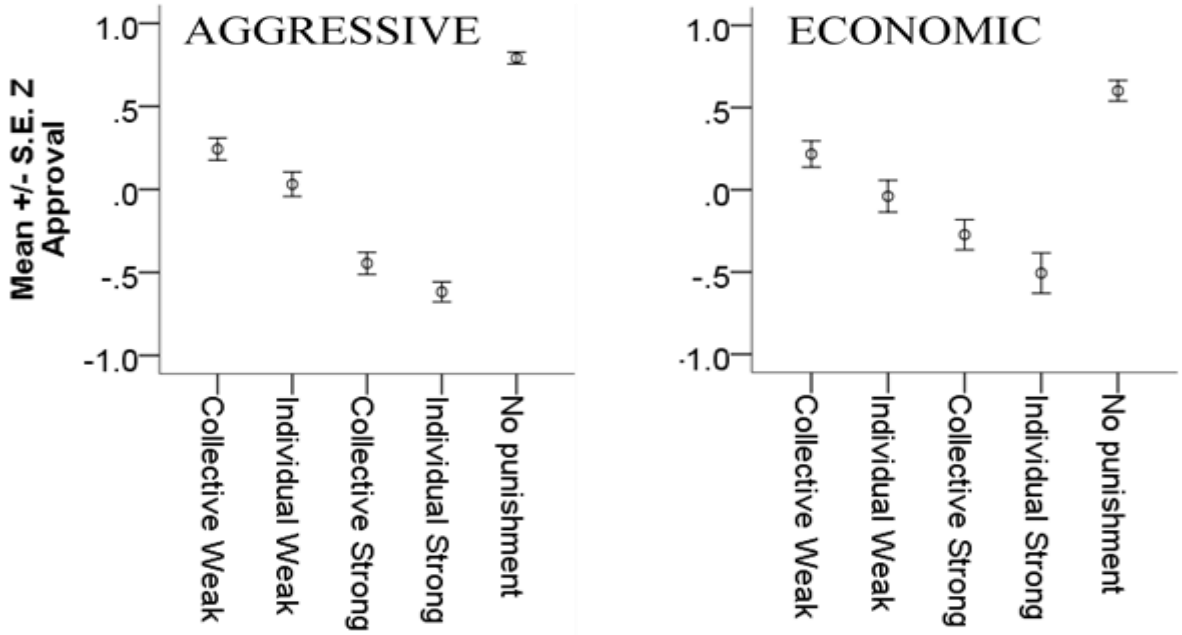


Figure 6. (Based on data from Eriksson et al., 2016). Mean punisher approval among US participants who watched animations where a selfish behavior was met with physical aggression (left) or economic punishment (right).

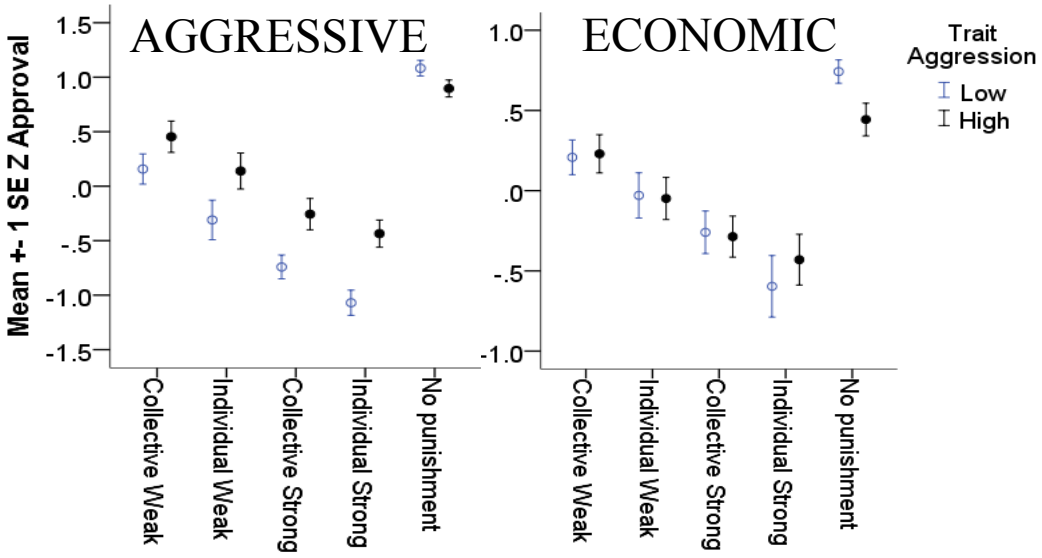


Figure 7. (Based on data from Eriksson et al., 2016). Mean punisher approval (as in Fig. 6) among participants that scored either low (blue bars) or high (black bars) on trait aggression according to a median split.

2.2.6 Conclusion

These studies set out to answer the question what social norms about punishment of selfish behavior look like. The gist of the answers we obtained is that these social norms seem to be mainly restrictive rather than prescriptive. To be expected to punish someone who behaves selfishly you must

have a very special role in relation to that person. Even then you can only choose punishment within a narrow range. Preferably, you should just explain to the selfish person that his or her behavior was wrong, ideally without even raising your voice and certainly not behaving aggressively.

When there are no special roles – a group of peers – our studies showed that individuals who take it upon themselves to punish a selfish group member tend to be judged as acting less appropriately than those who do not punish. The preferred alternative is for the group to manage punishment as a collective. Indeed, in her field studies of cooperation with respect to common pool resources, Ostrom (1990) found that individuals' responsibilities were typically restricted to monitoring whereas sanctions of non-cooperation were collectively managed.

A number of negative traits were also attributed to the punishers, such as being angry and likely to use punishment unfairly and create bad morale in the group. These attributions testify both to punishment being viewed as aggression and the negative consequences for the group that aggressive members are seen to have. Our findings of moderators of the disapproval of peer punishers give further support of the role of perceived aggressiveness. First, more aggressive raters tended to show less disapproval of peer punishment. Second, more severe punishments tended to be more disapproved of. Third, punishment that were directly aggressive but not harmful tended to be more disapproved of than punishment that were economically clearly harmful but not directly aggressive. It seems fair to conclude that dislike of aggressive group members is a major reason why use of confrontational peer punishment tends to have negative reputational effects.

These findings stand in stark contrast to the notion of punishment as a public good in itself. While peer punishment has the potential for benefitting the community by ensuring that people adhere to cooperative norms, it also has undeniable harmful effects. These harmful effects may subjectively overwhelm the beneficial effects. Whether they do so also objectively is another matter. Outside the laboratory it is unclear whether costs and benefits could even be measured, and extremely unclear whether they could then be compared in a meaningful way.

The results of our studies were remarkably constant across different scenarios and different countries. The possibility of human universals in norms regulating punishment is intriguing and warrants further study.

A concluding remark is that people do not disapprove of punishment in general. Political parties promising to enforce laws that punish wrongdoers tend to be popular. There is no popular demand for abolishing all punishment from the criminal code. In short, formal punishment is generally approved of. It is an important question what peer punishers can do to gain more approval for their actions.

2.3 Do people think using altruistic punishment is the moral thing to do to?

A third topic motivated by the altruistic punisher hypothesis is how people judge the moral status of costly punishment. The term "altruistic punishment" suggests it should be seen as a moral act. It is theoretically possible that costly peer punishment could still be seen a moral act even though we have found that it tends to be socially disapproved of. An example of the possibility of tension between morality and social approval is whistleblowing on corrupt practices in your organization, which is typically done out of moral duty in the face of harsh social sanctions from your colleagues (e.g., Dasgupta & Kesharwani, 2010). It is therefore relevant to study the morality of using costly

punishment in those economic games that have provided the main empirical basis of the altruistic punisher hypothesis: the public goods game and the ultimatum game.

In the literature on the public goods game, decisions to pay a cost to reduce a low-contributor's payoff have been conceived of as altruistic punishment. In the literature on the ultimatum game, decisions to reject an unfair offer have been conceived of as altruistic punishment. What the two decisions have in common is that they make both the decision-maker and another player (who has behaved selfishly) worse off. However, there are also notable differences between the decisions. One difference is who were harmed by the selfish behavior: in the public goods game it is an entire group, in the ultimatum game it is only the decision-maker. This is potentially important in that having an other-regarding disposition should be more important for the decision in the public goods game than in the ultimatum game. (But recall from Paper 1 that we did not find altruistic punishers in the public goods game to be very other-regarding either.)

Here we shall focus on another difference on which previous literature has been mute: the costly punishment decision is *framed* in different ways in the two games. In the public goods game the decision is explicitly framed as paying a cost to reduce the other's payoff (a "reduction" frame). In the ultimatum game the decision is instead framed as rejecting an unfair offer (a "rejection" frame). Indeed, the ultimatum game was originally introduced as a tool for the study of bargaining and the original paper made no mention of interpreting rejection as punishment (Güth et al., 1982). Thus, it is not clear that participants in the ultimatum game will process the decision about whether to reject as a decision about whether to punish. Participants could have other motives for using rejection, such as achieving fairness or asserting independence (Yamagishi et al., 2009).

In Paper 5 (Eriksson, Strimling, Andersson & Lindholm, 2017) we posed two questions and examined whether the answers depended on if costly punishment in the ultimatum game was framed as rejection or as reduction. First, to what extent is rejection/reduction seen as punishment? Second, to what extent is rejection/reduction seen as the moral choice?

2.3.1 To what extent is rejection/reduction seen as punishment?

Method. In Paper 5 (Exp. 4) we conducted a simple online survey in which an ultimatum game scenario was described, with two different versions of the last two sentences:

> Consider the following situation. Two people, anonymous to each other, take part in an on-line experiment. One of the two participants is named "proposer" and the other is named "receiver". The proposer is asked to make an offer about how to split 100 dollars between them, without any work required. This particular proposer's decision is to offer the receiver a share of 25 dollars (thus keeping 75 dollars for himself/herself).
>
> *Rejection version*: The receiver is now given the option to accept or reject this offer. If the offer is rejected neither participant receives any money.
>
> *Reduction version*: The receiver is now given the option to pay the 25 dollars he/she has earned in the experiment to reduce the other's payoff by three times as much. Thus, the proposer would then lose the 75 dollars he/she earned in the experiment.

The survey was taken by 200 US participants, half of which were assigned to each version. Participants in both conditions answered the same multiple-choice question on how they would interpret someone using rejection/reduction. There were four responses to choose from: the receiver is angry at the proposer and makes this decision to punish him/her (coded as Punishment); the receiver prefers a fair outcome to an unfair outcome (coded as Fairness); the receiver dislikes having the outcome decided by

another and wants to assert his/her independence (coded as Independence); the receiver has another motive than any of the above (coded as Other).

Results. The punishment interpretation was common for reduction (66%), but much less common for rejection (37%). See the left panel of Fig. 8.

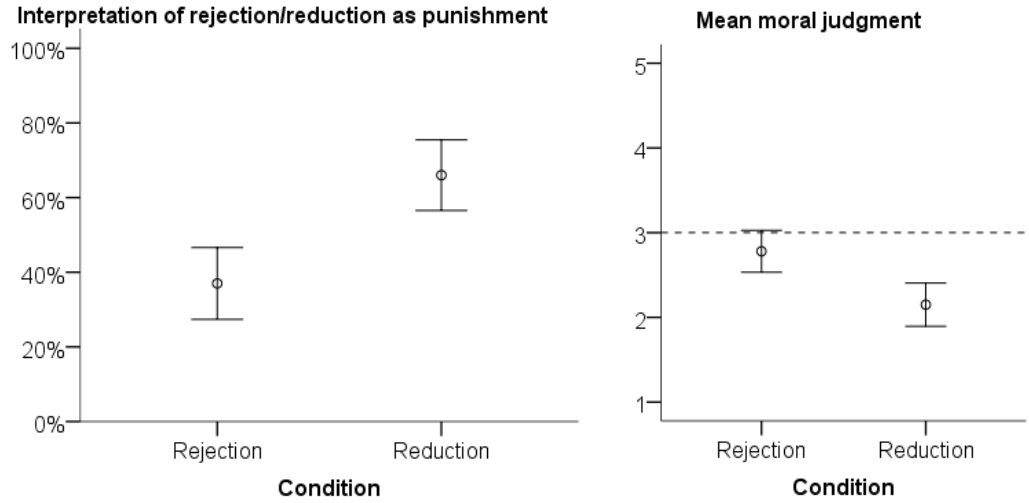


Figure 8. (Based on data from Eriksson et al., 2017). Proportion of participants who interpreted rejection/reduction as punishment (left), and mean moral judgment (right). Error bars indicate 95% confidence intervals. The dotted line indicate the neutral moral judgment.

2.3.2 To what extent is rejection/reduction seen as the moral choice?

Method. Our next study (Exp. 5) used the same stimuli and sample size as the previous one, but asked participants what would be the morally right thing for them to do if they were the second player in the scenario. Responses were given on a five-point scale, where 1 represented that use of rejection/reduction is definitely not the morally right thing to do, and 5 represented that use of rejection/reduction is definitely the morally right thing to do.

Results. The median moral judgment of rejection was 3, that is, neutral. The median moral judgment of reduction was 2, that is, not moral. See the right panel of Fig. 8.

2.3.3 To what extent are punishment or morality the motives for using and not using rejection/reduction?

Method. In Exp. 1-2 we examined the questions of how rejection and reduction relate to punishment and morality in another way: by having participants making decisions in the role of the second player and then asking them to motivate their decision. Exp. 1 used hypothetical decisions and a large sample (N=400). Exp. 2 replicated Exp. 1 with a smaller sample (N=168) in an incentivized actual game conducted online. The decision they had to make was the same as the one described in the scenarios in the previous studies, again framed as either rejection or reduction.

Results. As results were very similar between the hypothetical and incentivized experiment I here present the pooled data. The frequency with which a moral motive was invoked showed a remarkable pattern. Whereas nobody invoked a moral motive for using reduction, it was rather common (28%) to invoke a moral motive for *not* using reduction. In the rejection frame, the moral motive was essentially absent altogether. These results are consistent with the findings in the previous study of negative

moral judgments of reduction and neutral morals judgment of rejection. Consistent with moral concerns about reduction it was used substantially less than rejection, see the left panel of Fig. 9. A punitive motive was more common for using reduction than for using rejection, see the right panel of Fig. 9. This is consistent with the above finding that reduction is more likely to be interpreted as punishment.

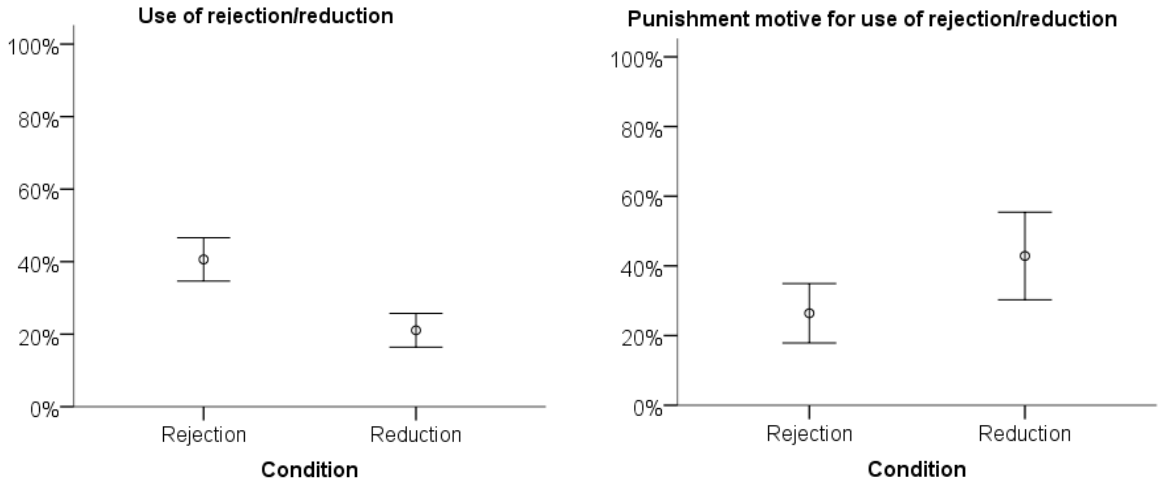


Figure 9. (Based on data from Eriksson et al., 2017). The proportion of participants who used rejection/reduction (left) and, among those who used it, the proportion who invoked a punitive motive (right). Error bars indicate 95% confidence intervals.

2.3.4 Conclusion

The results of these experiments indicate that the term altruistic punishment for rejection of unfair offers in the ultimatum game is a misnomer. For one thing, we would expect altruism to be associated with morality; however, we found little moral motivation for using rejection, and moral judgments of rejection were on average neutral. For another, the association between rejection and punishment is doubtful; rejection was typically not interpreted as punishment and the most common motive to use it was not to punish but to achieve a fair outcome. Indeed, it is a fallacy to assume that because an act has a certain consequence, that consequence was the reason for the act. Data on rejections in the ultimatum game may be a poor basis for psychological theories of punishment.

We also examined the same decision when it was framed explicitly as whether to reduce payoffs instead of whether to reject the unfair offer. In contrast to rejection, reduction was strongly associated with punishment. At the same time its association with moral behavior was negative. Thus, whereas rejection tended to be viewed neither as morally charged nor as punishment, reduction tended to be viewed as immoral punishment. This is not surprising from the perspective of moral psychology, as causing harm is a key trigger of negative moral judgment. Presumably, framing the decision as rejection instead of reduction makes the causing of harm less salient as it may now be perceived just as a side effect.

The framing effect was substantial also on behavior: use of rejection was almost twice as frequent as use of reduction, despite payoff consequences being identical. Such framing effects cannot be accounted for within the game theoretic paradigm, in which behavior is studied as an outcome of the game structure. Our finding is consistent with framing effects documented in previous research on the

ultimatum game (Larrick & Blount, 1997) and the dictator game (DeScioli & Krishna, 2013). What makes our finding stand out is that both the frames we used were already established as standard in the literature; we just transferred the reduction framing from the public goods game to the ultimatum game. For this reason, our finding calls into question whether it is at all worthwhile to strive for neutral language instead of purposefully inducing a frame (see also Eriksson & Strimling, 2014).

3 Theoretical integration

The empirical results I have summarized above strongly suggest that strong reciprocity theory paints an incorrect picture of the psychology of informal punishment of non-cooperators. This calls for an analysis of what is the fundamental problem with strong reciprocity theory. Here I will argue that the fundamental problem lies in the assumption that behaviors evolve specifically for cooperative situations (i.e., social dilemmas). I think it is more correct to assume that the psychology of informal punishment looks essentially the same across situations, regardless of whether or not they are of a social dilemma character. I conclude by sketching what a theory of informal punishment could look like and what kind of empirical questions such a theory would raise.

3.1 Where strong reciprocity theory goes wrong: the tacit assumption of game theoretic cognition

The general theory of evolution is based on replicators, that is, functional units that if they confer an advantage will spread in a population. Strong reciprocity theory is based on the notion that the strategy "cooperate and punish anyone who does not cooperate" is such a functional replicating unit. This notion makes perfect sense within the context of a game theoretic model of a social dilemma. But as humans do not live in the game theoretic model it is necessary to ask whether it makes sense to consider such a strategy as a functional replicating unit also in the real world of humans. A characteristic of the real world is that individuals encounter a myriad of situations, varying along many dimensions and none of them exactly identical to another. The above specification of the altruistic punisher strategy is therefore incomplete, because it also requires a preprocessing component for deciding whether the current situation is a social dilemma. I will refer to such a preprocessing component as *game theoretic cognition*. When game theoretic cognition is included in the specification the strategy could be expressed: "read off the current situation; if it is a social dilemma, cooperate and punish anyone who does not cooperate".

Evolutionary game theory models of animal behavior, such as birds fighting over a resource, do not rely on animals having game theoretic cognition but on recognition of a particular situation (e.g., that this is a fight over a resource) in which strategies are activated. Humans, however, encounter a very large malleable set of different situations and have a large and malleable repertoire of behavioral responses. For example, vaccination against measles can be seen as a social dilemma: it is very effective in preventing the disease but it comes with a little pain and takes a little effort, and if everyone else got vaccinated then you would be protected anyway. But vaccination has only been encountered in modern times, so it is unthinkable that humans would have evolved strategies specifically for this situation. Thus, strong reciprocity theory requires people to be equipped with the

game theoretic cognition to recognize vaccination as belonging to the category of situations in which altruistic punishment should be activated.

In a paper introducing functional interdependence theory, Balliet et al. (2016) noted that interdependence properties of situations (i.e., their game structure) are not readily observable and argued that humans should be equipped with adaptations for integrating relevant cues in order to produce accurate inferences about these situational properties. In my terminology, these authors argued that humans have game theoretic cognition. Their arguments centered on that such cognition would have been adaptive for our ancestors. However, it is not very meaningful to discuss the adaptiveness of a hypothetical trait. We need the trait in front of us to examine how it may be an adaptation. I think there is converging evidence for the *absence* of evolved modules amounting to game theoretic cognition. Specifically, it seems misleading to model people as wired with a set of strategies that are selectively activated on the basis of the game theoretic structure of the situations they encounter.

Some evidence come from cursory observations. For instance, if people had game theoretic cognition it seems weird that behavioral economists feel obliged to spend so much effort to make sure participants understand a game (e.g., by presenting instructions in writing as well as reading them out aloud, followed by various control questions). After all, compared to naturally occurring situations like vaccination, the experimental games are extremely unambiguous when it comes to the player set, the strategy set, and the payoffs. If people find it difficult to grasp even such explicitly presented games, how could they possibly be accurate at recognizing the game structure of naturally occurring situations? Another observation is that there seem to be no concepts in common language for abstract game theoretic structures such as social dilemmas. The term “free-riding” is sometimes used metaphorically (i.e., creating a correspondence between different situations by giving the same name to a behavior in one situation as to a corresponding behavior in another situation), but this usage seems to have originated among economists and is less than a hundred years old (http://conversableeconomist.blogspot.se/2014/10/more-on-origins-of-free-rider-idea.html). By contrast, all languages seem to have plenty of common words for personal qualities (e.g., courage, anger, wisdom, and power) and relations (e.g., family, friend, and foe) that transcend game theoretic structures.

More rigorous pieces of evidence come from papers in this dissertation. In Paper 1 we saw that the same individuals often use the same kind of costly punishment behavior across different situations: when someone made a low contribution, when someone made a high contribution, and when someone made contributions on the punisher’s behalf. Thus, use of costly punishment does not seem to be triggered by a specific game theoretic structure of the situation.

In Paper 5 we found that use of costly punishment in the ultimatum game depended on whether it was framed as rejection or reduction. This finding adds to a large body of research on framing that was pioneered by Tversky and Kahneman (1981). In brief, it is well-established that in situations where the game structure is made perfectly explicit, behavior is nonetheless very sensitive to cues that do not change the game structure.

Some unpublished data provide further evidence for the absence of game theoretic cognition. My colleagues and I have used online surveys to document that people do not see social dilemmas as

researchers do (de Barra, Strimling & Eriksson, submitted). Specifically, we presented participants with a number of situations that researchers generally consider to be social dilemmas (e.g., whether to use a face mask when you have a cold, whether to drive to work in your own car or using slower public transport or carpooling, and whether to join the army and protect the nation's security or leave it to others). The game theoretic structure of social dilemmas is such that you should prefer to defect regardless what others do, and you should prefer that others cooperate. However, that was not a common response pattern when we asked our participants for their preferences. Instead, the majority of respondents expressed a fixed preference for their own behavior – either to cooperate regardless of what others do or to defect regardless of what others do – and they preferred other people would behave the way they did, regardless if that was to cooperate or to defect. In other words, our data suggest that people do not evaluate behavior in terms of some objective payoffs to themselves and other people. Instead they tend to view the behavior they use themselves as morally superior.

In another working paper (Strimling, Bondesson & Eriksson, in preparation) my colleagues and I investigate social dilemmas and peer punishment in the earliest laws. Social dilemmas, in the sense of crimes committed against the group rather than against a specific individual, are almost completely absent in these laws. This is an indication that, historically, people have not easily recognized the importance of social dilemmas.

In sum, the evidence seem to be against the existence of game theoretic cognition, which in turn seems to invalidate strong reciprocity theory and similar evolutionary game theory based approaches to human cooperation.

3.2 A general theoretical model of disapproval and informal punishment

A key observation is that informal punishment is not used only in connection to social dilemmas; in fact, from everyday experience it seems that almost *any* behavior could be the object of informal punishment. In many cases this has been documented by researchers on social norms; let me just mention sociologist Norbert Elias and social psychologist Robert Cialdini as two researchers who have carried out influential empirical work on this topic. Here I will sketch a new theoretical model of disapproval and informal punishment, built on a few fundamental principles. First, the object of interest is an individual's processing of a situation and how it may result in a reaction that amounts to informal punishment (in contrast to a focus on evolved strategies or on social norms). Second, this processing and reaction can be analyzed as a sequence of judgments and decisions (each of which may be either automatic or deliberated on); this sequence is assumed to hold across individuals, but individual differences will play a crucial role for the results of these judgments and decisions. Third, the same framework should apply very generally. In particular, it should apply also to its own results, that is, the framework should describe also how someone's use of informal punishment is processed and reacted to by other actors. The model I propose involves a sequence of steps outlined below and is summarized in Fig. 10. It can be seen as an extension and refinement of the framework of Chaurand and Brauer (2008), which outlines three factors (hostile emotions, responsibility, and legitimacy) that determine whether someone speaks up against counter-normative behavior in urban environments (e.g., not picking up after one's dog).

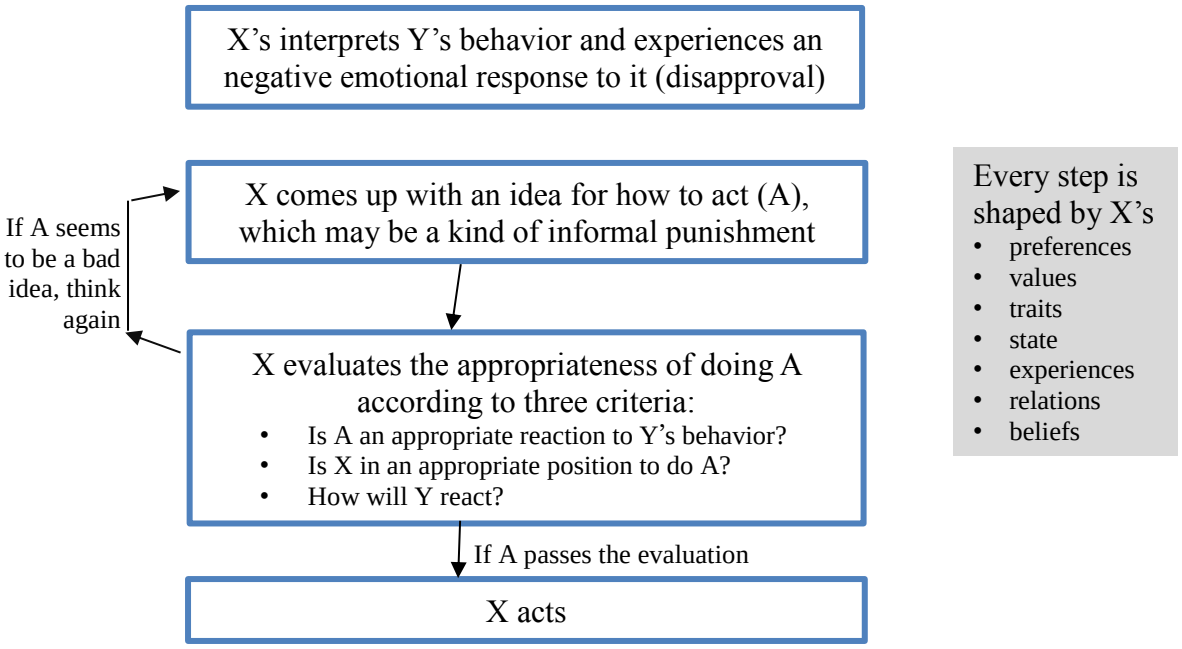


Figure 10. A model of disapproval and informal punishment.

3.2.1 Step 1: Interpretation of behavior and experience of emotional response

As a general scenario, consider agent X learning of agent Y doing behavior B (e.g., X and Y could be a married couple and the behavior could be putting the toothpaste tube back into its designated cup with the tube cap pointing down instead of up—something Y tends to do and X tends to dislike). X could learn of Y's behavior in various ways, including direct observation, someone else's account of what happened, and inference (say, from finding the toothpaste tube pointing the wrong way). In any case, X interprets the information to form a mental model of what happened (likely including the attribution of certain intentions and beliefs to Y) and experiences an emotional response to what happened. I think X's use of informal punishment crucially relies on the interpretation and emotional response of Y's behavior. In particular, I think it is extremely rare that people use informal punishment against a behavior they feel positive about, that is, approve of (whereas people with formal authority may be formally required to punish a behavior regardless of whether they approve of it), even if they know that many other people feel negative about it.

A key factor, for which individual variation is generally ignored in game theory based research, is *X's baseline preference for or against B*. People simply dislike different things. For example, this was evident in Paper 4, where we found less aggressive participants to be more disapproving of an agent who acted aggressively against a selfish agent (see Fig. 7). To understand patterns of use of informal punishment in a certain domain, I think the first thing to understand is how preferences are distributed.

Several additional factors will contribute to the emotional response, and there will be literatures on each of these factors to draw on. I do not delve into this literature here, as my aim is only to sketch the kind of framework I think would be useful for understanding informal punishment.

Some factors relate to Y, such as *X's attitude to Y*; if you already dislike someone you may disapprove of his or her behaving in a way that would be tolerable behavior from someone else. Another such factor is *X's attribution of Y's intent*; if you believe an otherwise bad behavior was for a good cause you may condone it.

Other factors have nothing to do with Y, such as *X's state*; if you are already irritated, stressed, hungry, in pain, etc., you may disapprove of a behavior that you generally tolerate. Another factor is *X's judgment of the consequences of this instance of B*. The same behavior may be seen to have worse consequences in a certain context and therefore earn more disapproval. An extreme case of this was displayed in Paper 1, where we changed the public goods game such that the negative consequences of the act of "contributing to the common pot" moved from the contributor to another player, who then tended to disapprove more of such contributions.

3.2.2 Step 2: Come up with an idea for how to act

Assuming that X disapproves sufficiently of Y's behavior to do something about it, there are generally a wide range of options available (in contrast to economic experiments, where there is usually only the choice of whether to reduce payoffs). The following list covers some common options.

- *Demand remorse, and possibly compensation, from Y*. The kind of punishment that is available in economic experiments amounts to a fine, i.e., to removing some resource from the other. Papers 3 and 4 indicate that it may not be an acceptable reaction. In real-life situations I think it is not very common that the option of a fine is available, but very common that you demand an apology and perhaps a compensation, either substantial or symbolic.
- *Lash out physically at Y*. Physical aggression was typically rated as an unacceptable reaction to selfish behavior in our studies in Papers 3 and 4. It is likely to be more acceptable as a reaction to another's physical aggression.
- *Show Y that you are upset*. An angry reprimand or a display of sadness are reactions that are almost always available and used in many situations.
- *Anonymously show Y that someone is upset*. This could be an unsigned message, a poster board note, etc.
- *Reason with Y*. In Paper 3 we found that the most acceptable reaction tend to be to just explain why you think a behavior is wrong.
- *Avoid Y*. This could mean averting your eyes, not talking to Y, walking away, etc.
- *Inform some authority*. Depending on the situation, there may be some parent, teacher, boss, or other person in a supervisory position you could alert. This may lead to that person taking action, but also to that person changing their attitude to Y.
- *Engage an appropriate group to take collective action with respect to Y*. In Papers 3 and 4 we found that use of informal punishment may be more acceptable if an appropriate group is collectively engaged.
- *Inform someone else*. Many people will gossip about what someone did. This may be a way of dealing with one's emotions and possibly to make sure that someone else knows what happened. It may also lead to that person giving you advice on how to act, or even acting on their own.

- *Make joking comment to Y.* Although not studied in this dissertation, I think joking is a common response to behavior you disapprove of. A joke may indicate your disapproval while maintaining a friendly atmosphere.
- *Support Y.* Although not studied in this dissertation, one may react to someone's bad behavior by a positive response (like a hug) as a way to improve relationships and make the other feel better.
- *Change your own behavior in order to give Y less reason to continue with his/her behavior.* Perhaps Y's behavior should be interpreted as an informal punishment of you!
- *Change your own behavior in order to make Y's behavior more difficult.* Perhaps Y's behavior should be interpreted as an informal punishment of you!
- *Remember what Y did.* It may affect how you act and feel next time Y does something.
- *Do nothing*. There are so many little nuisances in the world, you cannot take action every time.

Which action first comes to mind in a situation may depend on situational constraints, but also on individual differences. It seems likely that some people, due to some combination of personality and life history, are habitual avoiders whereas others are habitual reasoners, verbal punishers, jokers, etc.

3.2.3 Step 3: Evaluation of the appropriateness of the suggestion action

An impulsive individual might act according to the first idea that comes to mind without a second thought. However, I think actions that amount to informal punishment are typically influenced by various considerations that can be organized into three themes:

Is the action an appropriate reaction to Y's behavior? As discussed above, the appropriateness of an action will depend on what it is a reaction to. In a paper that just came out, we found that a fixed reaction (a reprimand) could be deemed as anything from strongly inappropriate to strongly appropriate depending on the severity of the norm violation it was a reaction to (Eriksson, Andersson, & Strimling, 2017). In the same paper, we found that reprimands were deemed as less appropriate if anger was shown—unless the norm violation was severe, in which case showing anger had no effect on appropriateness ratings.

Am I in an appropriate position to take this action? A key finding in Papers 2 and 3 was that people care a lot about *who* uses informal punishment in a given situation. Specifically, it seems that closeness, superiority, and being directly affect by the other's behavior, give you more right and obligation to use informal punishment. Moreover, you seem to have less right to use informal punishment if someone else is closer, or more superior, or is more affected by the behavior than you are. Even the presence of other individuals that are indistinguishable from you in these respects (i.e., peers) seemed to make it less appropriate for you to use informal punishment. I think the positional aspect is a very important factor in who uses informal punishment in any given situation. (It is also typically neglected in game theory based accounts.)

How will Y react? A strong reason for not using informal punishment is that it may lead Y to retaliate instead of changing to a more approved behavior. Indeed, recall that Nikiforakis (2008) showed in an economic experiment how the presence of an option for counter-punishment made the

punishment frequency radically decrease (such that the positive effect of punishment on behavior was eliminated). To predict what Y's reaction is likely to be, people may use a number of cues including whether Y is likely to see your disapproval as legitimate (rather than as your pet peeve), the proportionality of your action to Y's behavior, your relation to Y and Y's relation to others in the situation, Y's personality and capability of retaliation, etc. Of course, there will also be individual differences in how much you care about Y's reaction.

3.3 Directions for future research

In this final section I briefly discuss a number of research questions that arises from or are informed by the model of informal punishment presented above. I am already working on some of these questions.

What is the scope of the model? My model is a radical departure from game theory based theories of informal punishment, such as strong reciprocity theory and much other work on social norms that takes from granted that norms develop specifically around social dilemmas. My model is much more in line with the social psychological work of Chaurand and Brauer (2008) on determinants of social control of uncivil behavior. However, it extends that work in several ways. A key extension is my assumption that the target behavior does not need to be generally considered as uncivil; rather, I explicitly assume that the same process applies whatever the target behavior. I also assume that people have very wide-reaching and varying preferences for others' behavior, which would mean that my model has a very large scope. The validity of these assumptions is an important question for future research. It ties into several of the other questions below.

Where do preferences for others' behavior come from? If I am right that almost no behavior is completely shielded from causing disapproval in someone, how can that be? A core explanation, I believe, is that people have preferences for their *own* behavior (which is less puzzling), and that a feature of human psychology is that people tend to automatically translate these preferences to apply also to others' behavior. This is the topic of current research in my group.

What is uniquely human about informal punishment? Recall the "evolutionary puzzle of human cooperation" according to strong reciprocity theory (section 1.3). To the extent that informal punishment is a key to human cooperation, it should be interesting to examine what is uniquely human in my model. I believe that a uniquely human feature is exactly the capacity for assigning approval or disapproval to any possible behavior. Many animals punish others for infringing on their territory or on their resources, but they seem to have little capacity for caring about behavior that does not directly affect themselves, such as whether another does something good for the group.

How universal are the rules of appropriateness? Papers 3 and 4 identified a couple of rules for the appropriateness of informal punishment that seem to hold in some generality. Specifically, greater severity tends to make informal punishment less appropriate whereas collective involvement tends to make informal punishment more appropriate. A former postdoc of mine just published a study that I helped design, which found that the same rules seem to apply also among the Turkana, a pastoralist tribe in Kenya (Mathew, 2017). Moreover, the same rules applied both in a situation of a social dilemma character and in a situation that is not a social dilemma. I also just completed a study where the geometric animations from Paper 4 were used to measure social perceptions of peer punishers in

eight countries (Eriksson et al., in press). Collectively managed peer punishment gained higher ratings than individually managed peer punishment in each of eight countries included in the study (Sweden, Netherlands, United States, Japan, China, Russia, United Arab Emirates, Pakistan). So far, I know of no cultures where these rules of appropriateness do not hold.

How do cultures differ in the use and appropriateness of informal punishment? Although my basic model of informal punishment is assumed to hold very generally, a key ingredient is the role of individual differences and they may translate to cultural differences. There are obviously cultural differences in which behaviors tend to be punished, often thought of as cultural norms of behavior. There may also be cultural differences in which informal punishment is used. In the abovementioned study (Eriksson et al., in press), we found that peer punishers using aggressive confrontation were clearly disapproved of in more individualistic countries (Sweden, Netherlands, United States) but not disapproved of in less individualistic countries (China, Russia, United Arab Emirates, Pakistan), with Japan at an intermediate level. My interpretation of these findings is that the extent to which confrontational punishment is condoned may depend on a culture's emphasis on individuals' rights, as such rights may include the right to break a social norm without being punished for it. The interaction between confrontational punishment and culture is important, not least because social norms seem to be always gradually changing and confrontational punishment is a mechanism for sustaining them. Cultural differences in the support of confrontational punishment may therefore influence the speed by which social norms change in different cultures. I am currently applying for grants to pursue this line of research together with several international partners.

Can we understand why targets of informal punishment change over time? My main research interest lies in cultural evolution. A particularly interesting question in this field is why norms of behavior change over time. This question is under-explored. Indeed, many theoretical treatments of norms take for granted that they do not change (e.g., by conceiving of a norm as an "equilibrium"). However, it is well-established that norms about hygiene and violence have exhibited mainly unidirectional change over a long period of time and in recent work (Strimling, de Barra & Eriksson, submitted) we propose an explanation for this change based on the notion of systematic differences in use of informal punishment. First, in line with my model, we show that both those who prefer more hygienic (or less violent) behavior and those who prefer the opposite are willing to use informal punishment against the "other side". Second, we show that those who prefer more hygienic (or less violent) behavior are *more* likely than those who prefer the opposite to use it. This asymmetry could drive norm change in one direction. This adds a new tool for those who study cultural evolution: where directional cultural change occurs, look for a fundamental asymmetry at the individual level. In an ongoing project we use the same tool to explain liberalization of moral opinions.

To conclude, although the research presented in this dissertation was mainly motivated by skepticism of an influential theory of punishment, I believe it also serves as a basis for theoretical developments that have the potential of guiding empirical research in several novel directions.

References

Argyle, M. (1991). *Cooperation: The Basis of Sociability*. Routledge, New York.

Balliet, D., Tybur, J. M., & Van Lange, P. A. (2016). Functional interdependence theory: An evolutionary account of social situations. *Personality and Social Psychology Review*. DOI: 10.1177/1088868316657965.

Binmore, K., & Shaked, A. (2010). Experimental economics: Where next?. *Journal of Economic Behavior & Organization*, *73*(1), 87-100.

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. (2003) The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100* (6), 3531–3535.

Bryant, F. B., & Smith, B. D. (2001). Refining the architecture of aggression: A measurement model for the Buss–Perry aggression questionnaire. *Journal of Research in Personality, 35*, 138–167.

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.

Cownden, D., Eriksson, K., & Strimling, P. (2017). A popular misapplication of evolutionary modeling to the study of human cooperation. *Evolution and Human Behavior*, *38,* 421–427.

Dasgupta, S., & Kesharwani, A. (2010). Whistleblowing: a survey of literature. *IUP Journal of Corporate Governance, 9*(4), 57–70.

DeScioli, P., & Krishna, S. (2013). Giving to whom? Altruism in different types of relationships. *Journal of Economic Psychology, 34*, 218–228.

de Barra, M., Strimling, P., & Eriksson, K. (submitted). It's the wrong game! The strategic structure of hygiene-related behaviors.

Eriksson, K., Andersson, P.A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes and Intergroup Relations*, 19(2), 152–168.

Eriksson, K., Andersson, P.A., & Strimling, P. (2017). When is it appropriate to reprimand a norm violation? The roles of anger, behavioral consequences, violation severity, and social distance. *Judgment and Decision Making, 12,* 396–407.

Eriksson, K., Cownden, D., Ehn, M., & Strimling, P. (2014). 'Altruistic' and 'antisocial ' punishers are one and the same. *Review of Behavioral Economics, 1,* 1–13.

Eriksson, K., & Strimling, P. (2009). Partner search heuristics in the lab: Stability of matchings under various preference structures, *Adaptive Behavior*, *17*(6),524–536.

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

Eriksson, K., & Strimling, P. (2014). Spontaneous associations and label framing have similar effects in the public goods game. *Judgment and Decision Making, 9*, 360–372.

Eriksson, K., Strimling, P., & Ehn, M. (2013). Ubiquity and efficiency of restrictions on informal punishment rights. *Journal of Evolutionary Psychology, 11,* 17–34.

Eriksson, K., Strimling, P, Andersson, P.A., Aveyard, M., Brauer, M., Gritskov, V., Kiyonari, T., ... (in press). Cultural universals and cultural variation in meta-norms about peer punishment. To appear in *Management and Organization Review*.

Eriksson, K., Strimling, P., Andersson P.A., & Lindholm, T. (2017). Costly punishment in the ultimatum game evokes moral concern, in particular when framed as payoff reduction. *Journal of Experimental Social Psychology*, *69*, 59–64.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*, 785–791.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*(1), 1–25.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*(4), 980–994.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. The *Quarterly Journal of Economics, 114*(3), 817–868.

Fiske, A. P. (1992). The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological Review, 99*(4), 689–723.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., ... & Aycan, Z. (2011). Differences between tight and loose cultures: A 33-nation study. *Science, 332*(6033), 1100–1104.

Greitemeyer, T., Fischer, P., Kastenmüller, A., & Frey, D. (2006). Civil courage and helping behavior: Differences and similarities. *European Psychologist, 11*, 90–98.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3,* 367–388.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57*(2), 243–259.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*(1), 79–89.

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

Kahneman, D., Schkade, D., & Sunstein, C. (1998). Shared outrage and erratic awards: The psychology of punitive damages. *Journal of Risk and Uncertainty, 16*(1), 49–86.

Kiyonari, T., & 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,* 826–842.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*, 75–84.

Larrick, R. P., & Blount, S. (1997). The claiming effect: Why players are more generous in social dilemmas than in ultimatum games. *Journal of Personality and Social Psychology, 72*(4), 810–825.

Mathew, S. (2017). How the second-order free rider problem is solved in a small-scale society. *American Economic Review, 107*, 578–581

Messick, D. M, & McClintock, C. G. (1968). Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology, 4*, 1–25.

Nakao, H., & Machery, E. (2012). The evolution of punishment. *Biology & Philosophy, 27,* 833–850.

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.

Ostrom, Elinor (1990), *Governing the Commons*. New York: Cambridge University Press.

Strimling, P., de Barra, M. & Eriksson, K. (submitted). The civilizing process revisited: How asymmetries in punishment propensity may drive norm change.

Strimling, P., Bondesson, M. W. & Eriksson, K. (in preparation). Were social dilemmas and peer punishment important in pre-historic societies? Answers from a study of early laws.

Strimling, P., & Eriksson, K. (2014). Regulating the regulation: Norms about how people may punish each other. In van Lange, P., Yamagishi, T., Rockenbach, B. (eds.), *Reward and Punishment in Social Dilemmas*. Oxford University Press, pp. 52–69.

Tversky, A., and Kahneman, D. 1981. The framing of decisions and the psychology of choice. *Science, 21,* 453–458.

Van Lange, P. A. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology, 77*(2), 337–349.

Van Lange, P. A., Rockenbach, B., & Yamagishi, T. (Eds.). (2014). Reward and punishment in social dilemmas. Oxford University Press.

Wenzel, M., & Okimoto, T. G. (2016). Retributive justice. In *Handbook of Social Justice Theory and Research* (pp. 237–256). Springer New York.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*(1), 110–116.

Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., & Cook, K. S. (2009). The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Sciences, 106,* 11520–11523.

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., ... Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences, 109*, 20364–20368.