THE IMPORTANCE OF EMOTIONS FOR THE EFFECTIVENESS OF SOCIAL PUNISHMENT[*]

ASTRID HOPFENSITZ[§]

ERNESTO REUBEN[∝]

ABSTRACT: This paper experimentally explores how the enforcement of cooperative behavior in a social dilemma is facilitated through institutional as well as emotional mechanisms. Recent studies emphasize the importance of negatively valued emotions, such as anger, which motivate individuals to punish free riders. However, these types of emotions also trigger retaliatory behavior by the punished individuals. This makes the enforcement of a cooperative norm more costly. We show that in addition to anger, 'social' emotions like shame and guilt need to be present for punishment to be an effective deterrent of uncooperative actions. They play a key role by subduing the desire of punished individuals to retaliate and by motivating them to behave more cooperatively in the future.

July 2005

JEL Codes: Z13, C92, D74, H41

[*] We would like to thank Sam Bowles, Arno Riedl and Frans van Winden for useful comments and suggestions.

[§] University of Amsterdam and Tinbergen Institute. Email: a.hopfensitz@uva.nl.

[∝] University of Amsterdam and Tinbergen Institute. Email: e.reuben@uva.nl.

1. Introduction

Cooperation in social dilemmas is a phenomenon that is hard to explain but important to understand. Contrary to the predictions of theories that assume rational own-payoff-maximizing individuals, people cooperate with each other in many situations (e.g. Ostrom, 1998). The existence and enforcement of social norms seem to be an important mechanism for the promotion of cooperation. As shown by Fehr and Gächter (2000), cooperative behavior can persist when there is an opportunity to punish defectors. However, although punishment can have desirable consequences, it can also have a negative effect on welfare (Fehr and Rockenbach, 2003; Egas and Riedl, 2005; Gächter and Herrmann, 2005). To correctly predict when punishment will have positive results, we must understand the behavior of individuals who punish as well as that of individuals who are punished. To do this, one must realize that emotions play an important role in decision-making (Damasio, 1994; Loewenstein, 1996; Elster, 1999; Thaler, 2000).

The goal of this paper is to understand the motivations behind the behavior of both the punishers and the punished, and in particular, the type of motivations that must be present for punishment to be an effective institution for the promotion of cooperation. We concentrate on the role of social emotions, such as shame and guilt, as an essential component for the successful enforcement of cooperative norms.

Recent research has revealed that emotions motivate individuals to punish opportunistic behavior. In particular, anger has been shown to be of influence when subjects have to decide whether to punish or not. Unkind behavior induces anger and the angrier people are, the more likely they are to incur costs in order to penalize such behavior (Bosman and van Winden, 2002; Quervain et al., 2004). But anger cannot explain why punishment is actually effective. The effectiveness of punishment depends on the reaction of the individuals who are punished. If individuals feel anger after being punished, they may be motivated to retaliate towards the punisher. Therefore, anger alone may induce multiple rounds of punishment and retaliation and consequently a significant destruction of resources. What is missing to make punishment effective is a moral reaction of the punished. That is, after receiving punishment the punished should act more cooperatively and abstain from retaliation. We will show that the social emotions of shame and guilt motivate individuals to react in precisely this way.

Moral behavior has been shown to be critically linked to the ability for emotional reactions (Anderson et al., 1999; Moll et al., 2002). While this is true for emotional reactivity in general, of particular importance are emotions that facilitate prosocial behavior (i.e. prosocial emotions such as shame, guilt and empathy; see Bowles and Gintis, 2001). They do so by inducing a feeling of discomfort when doing something that violates one's values or norms, or those of other agents whose opinion one cares about. Shame and guilt are both ‘self-reproach’ emotions elicited by the individuals’ own blameworthy actions (Ortony, Clore, and Collins, 1988). While they differ in multiple dimensions concerning elicitation and action tendency (Tangney and Dearing, 2002), they are two very similar emotions and are often elicited at the same time.

The influence of prosocial emotions on behavior might be twofold. First, the anticipation and wish of avoidance of shame and guilt might induce norm-abiding behavior. Second, the experience of shame or guilt, after an action, might lead to behaviors to diminish the feeling. This can be through repayment, future cooperation or avoidance of contact with the interaction partner. If these emotions are elicited through punishment of selfish behavior, they might inhibit retaliation and encourage individuals to act more cooperatively in the future.

To test whether this true, we study, by means of an experiment, cooperation and punishment behavior in a social dilemma game. We introduce a more realistic form of social punishment where individuals who are punished always have the opportunity to retaliate. After all, if there is access to a punishment technology, it is likely that both the punisher and the punished have access to it. Indeed, we find that many individuals punish back after being punished. In various cases this escalates as individuals punish each other in turns, resulting in considerable welfare losses. Nevertheless, this punishment institution is still effective for sustaining cooperation.

In order to explain the behavior of both punishers and punished, we control for the emotional experience of ‘punishment-inducing’ emotions such as anger and irritation and ‘norm-enforcing’ emotions like shame and guilt. An important finding is that individuals that act unkindly do nevertheless feel considerably angry when punished. Consequently, punishment advances cooperation only when feelings of shame restrain the anger-induced desire to fight back. Finally, in order to observe the effect of punishment on future cooperative behavior, we had individuals play the game twice. We find that individuals are more likely to act kindly in the future only when punishment induces feelings of shame.

The paper is organized as follows. In Section 2 we describe the design of the experiment. Section 3 describes the subjects' behavior. In Sections 4 and 5 we analyze the relationship between the emotions and the behavior of the punishers and the punished. Section 6 discusses the results and concludes.

2. The Experiment

Lately, punishment mechanisms are analyzed in the context of public good games (using the framework of Fehr and Gächter, 2000). However, in this paper we require a simpler setting where the causes and effects of emotions can be easily observed and analyzed. To study the impact of social emotions, we used a two-person social dilemma game with and without punishment opportunities. Our game is similar to many of the social dilemma games in the literature, such as, the sequential prisoners' dilemma, the investment game, the trust game, etc.

2.1. The game

We first describe the game without punishment opportunities and then we explain how punishment is introduced. The game consists of two players taking part in a one-shot game. We will refer to these players as the 'first mover' and the 'second mover'. At the start of the game, the first mover receives 150 points whereas the second mover receives 100 points (see Figure 1 for the game tree). In the first stage, the first mover decides to either defect or cooperate. If the first mover defects, he keeps his 150 points, the second mover keeps her 100 points, and the game ends. If the first mover cooperates, 50 of his 150 points are multiplied by six and transferred to the second mover. Thus the second mover receives 300 points while the first mover loses 50 points. In the second stage, the second mover returns an amount of points (r) back to the first mover. Specifically, she could return 150 points (an equal split of the gains), 50 points (returning exactly the points lost by the first mover) or 0 points. After the decision of the second mover the game ends. Hence, if the first mover cooperates his payoff is $\pi_1 = 100 + r$ and the payoff of the second mover is $\pi_2 = 100 + 6 \times 50 - r$. This describes the game without punishment.

In the game with punishment both players can assign punishment points. Doing so is costly for both players. We denote p_{it} as the amount of points assigned by player i (for $i \in \{1,2\}$) in punishment round t. After the second mover decides how much to return, the first round of punishment starts. First, the first mover gets the opportunity to assign a nonnegative amount of punishment points to the second mover (p_{11}). The first mover looses p_{11} points and the second mover looses $4 \times p_{11}$ points. In order to avoid large losses during the experiment, the first mover could assign punishment points only as long as the second mover had a positive number of points (i.e. $\frac{1}{4}(100 + 6 \times 50 - r) \geq p_{11} \geq 0$). If the first mover chooses $p_{11} = 0$ the game ends. However, if the first mover chooses $p_{11} > 0$ the second mover gets the opportunity to assign punishment points to the first mover (p_{21}). In order to avoid confusion, we will refer to punishment by the second mover as retaliation. Punishment by first movers and retaliation by second movers had the same cost and did the same harm. Thus for each retaliation point assigned, the first mover looses four points. Like the first mover, the second mover could assign retaliation points only as long the first mover had a positive number of points (i.e. $\frac{1}{4}(100 + \mathrm{r} - p_{11}) \geq p_{21} \geq 0$). If $p_{21} = 0$ the game ends, but if $p_{21} > 0$ the game continues with a second round of punishment. That is, the first mover gets the opportunity to assign additional punishment points to the second mover (p_{12}). Again, if $p_{12} = 0$ the game ends but if $p_{12} > 0$, the second mover gets the opportunity to assign additional retaliation points (p_{22}), and so on. The process repeats itself until either one of the players has zero points and therefore can not be punished further, or one of the players decides to assign zero punishment points. Therefore, if the first mover cooperates his payoff is $\pi_1 = 100 + r - \Sigma_t p_{1t} - 4 \times \Sigma_t p_{2t}$ and the payoff of the second mover is $\pi_2 = 100 + 6 \times 50 - r - \Sigma_t p_{2t} - 4 \times \Sigma_t p_{1t}$.

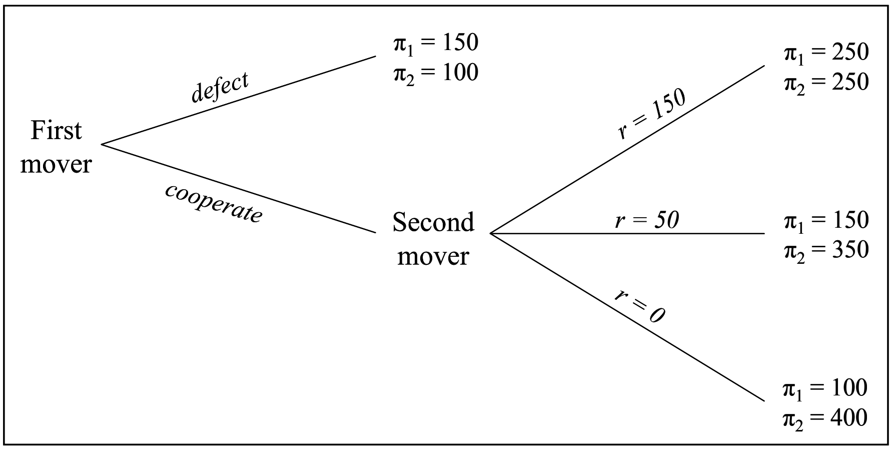

FIGURE 1 – GAME TREE IN THE CASE OF NO PUNISHMENT OPPORTUNITIES

If we use the standard assumption of rational individuals with self-regarding preferences, the unique subgame perfect Nash equilibrium of the game with and without punishment, is for second movers to return zero points and thus for first movers not to cooperate.[1] The predictions can change if individuals possess other-regarding preferences such as a concern for unequal payoffs, efficient outcomes, and/or reciprocating kind and unkind actions.[2] In the game without punishment, if the frequency of selfish individuals is sufficiently low then there can be equilibria where some second movers return positive amounts and some first movers cooperate. In the game with punishment, in addition to individuals who are willing to act kindly, there might be individuals who are willing to punish selfish behavior. If punishment leads to higher returns from the second movers, it gives first movers a further incentive to cooperate.[3] Certainly, the first movers' willingness to punish depends on the willingness of second movers to retaliate, which in turn depends on the willingness of first movers to punish once again, and so on. This, in our opinion is a more realistic way of modeling social punishment. That is, both the punisher and the punished have access to the punishment technology, and hence the punished always get the opportunity to retaliate. Moreover, both players always have the option to avoid further interaction by deciding not to punish. To our knowledge there is no other paper which examines the punishment behavior of individuals in such a setting.[4]

2.2. Experimental design and procedures

The computerized experiment was conducted in March 2005 in the CREED laboratory at the University of Amsterdam. In total 162 students from the University of Amsterdam participated in the experiment. Approximately 54% were students of economics and the rest came from a variety of fields such as biology, political science, law, and psychology. The average age was 22 years and 58% of the participants were male.

[1] Note that since punishment is always costly, it is never credible at any stage.

[2] See Rabin (1993), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Falk and Fischbacher (2000), Charness and Rabin (2002), and Dufwenberg and Kirchsteiger (2005).

[3] For example, using the same assumptions they use about the distribution of types, the model of Fehr and Schmidt (1999) predicts that, in the case of no punishment, 40% of second movers would return 150 points. However, this is not enough to instigate first movers to cooperate. Therefore cooperation is nonexistent. In the case of punishment, there are enough first movers that would punish so that all second movers return 150 points and hence all first movers cooperate.

[4] Nikiforakis (2004) studies punishment in a public good game in which retaliation was possible. However, in this case the punishment phase automatically ended after retaliation. As we will see, this restriction might have limited the amount of initial punishment.

Each subject played *twice* the social dilemma game described in the previous section. We used a perfect strangers matching protocol to avoid any reputation effects. In total, 26 subjects participated in the baseline treatment, that is, the game without punishment opportunities. The remaining 136 subjects participated in the punishment treatment. Earnings were calculated in points and points were exchanged for money at a rate of 40 points for 1 euro. The average earnings were 10.55 euros (this includes a show-up fee of 1 euro). The whole experiment lasted about one hour. Subjects were recruited through the CREED recruitment website and the experiment was programmed with z-Tree (Fischbacher, 1999).

After arrival in the reception room, subjects were randomly assigned to a table in the lab. Once everyone was seated, subjects were given the instructions for the experiment (see Appendix A). Subjects were told that the experiment consisted of two independent parts. We emphasized the fact that they will interact with different individuals in each part, and that, their choices in the first part would not affect their earnings in the second part. After this, the one-shot social dilemma game was described as the first part of the experiment. When everybody had finished reading the instructions, subjects had to answer a few questions to ensure their understanding of the game. Subsequently, the subjects played the social dilemma game via the computer (part 1). At the end of the first part, instructions were distributed concerning the second part of the experiment. The instructions informed subjects that they were about to play the same game once again. Furthermore, they would be in the same position as in the first part (i.e. first or second mover), and with certainty, their partner would not be the same partner they had played with in the first part. After they played the second part of the experiment (part 2), subjects filled in a debriefing questionnaire and thereafter they were paid out their earnings in private and dismissed.

To observe if emotional reactions of shame and guilt influence behavior, we used self reports to measure these and other emotions during the game. We also measured expectations concerning the behavior of the other player and fairness perceptions. Emotions were always measured after subjects observed the choice of the other player but before they made their own choice. Expectations about the behavior of the other player were asked after the subjects made their choice but before they observed the other player’s actual choice. Finally, fairness perceptions were measured at the end of the experiment in the debriefing questionnaire.

We measured emotions through self-reports, which are a reliable and often used technique in social psychology (Robinson and Clore, 2002) and have been shown to be correlated with physiological measures of arousal (Ben-Shakhar et al., 2004). We also used self-reported measures of expectations and fairness perceptions. Emotions and fairness perceptions were measured using seven-point scales, and expectations were measured by asking for a point estimate of the most likely action.[5] We measured a variety of emotions to avoid pushing subjects in a particular direction. The measured emotions were: anger, gratitude, guilt, happiness, irritation, shame, and surprise.

3. Observed Behavior

In this section, we give an overview and a brief discussion of the behavior of first and second movers. A summary of the behavioral data can be found in Appendix B. We start by investigating how often first movers cooperate and, when given the opportunity, how much second movers return. Comparing the baseline and the punishment treatments allows us to observe the effect of the punishment institution on the subjects' behavior. Then, in order to explain any differences induced by punishment, we analyze the punishment behavior of first movers as well as the retaliatory behavior of second movers. Finally, we examine whether the opportunity to punish has an effect on how subjects adjust their behavior from part 1 to part 2.

3.1. Cooperation and Returns

Figure 2 summarizes the main differences between the baseline and the punishment treatment. Namely, first movers cooperate more often and second movers return more in the presence of punishment.

As can be seen in Figure 2A, in both treatments, almost all first movers cooperate in the first part (more than 84.6%). However, in the absence of punishment, cooperation decreases substantially in the second part. If there are punishment opportunities, first movers cooperate equally often in both parts. Testing for differences between treatments confirms this observation. There is no significant

[5] Emotional intensity was measured from: 1 = 'not at all' to 7 = 'very intensely'. The fairness of an action was measured from: 1 = 'very unfair' to 7 = 'very fair'. The questions used are available in Appendix A.

difference in the frequency of cooperation in the first part (p = 0.837) but a highly significant difference in the second (p < 0.001).[6] There is an even starker difference between treatments when we consider the behavior of second movers. That is, in each part, second movers return noticeably less in the absence of punishment (p < 0.044). Given this behavior of second movers, it is easy to understand the decrease in cooperation in the baseline treatment. Remember that first movers who cooperate send 50 points. In the baseline treatment, they receive on average a smaller amount in return. In contrast, first movers who cooperate in the punishment treatment receive back roughly twice the sent amount. It is clear that, even when it is possible to retaliate, punishment limits the opportunistic behavior of second movers. In the following paragraphs, we examine how subjects punish and retaliate.

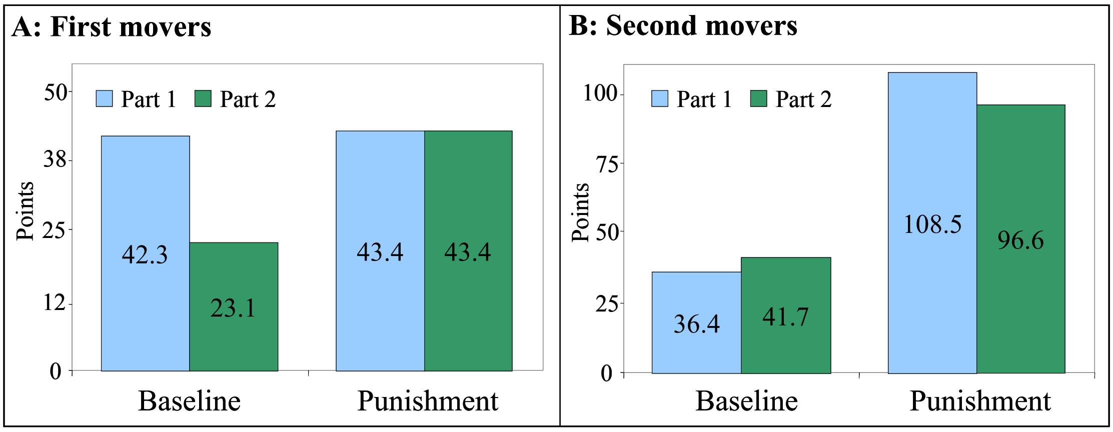

FIGURE 2 – COOPERATION BY FIRST MOVERS AND RETURNS BY SECOND MOVERS

Note: A) Mean number of points sent by first movers in each part and treatment. Note that, since first movers could send only 0 or 50 points, this is equivalent to the frequency of first movers who cooperate. B) Mean number of points returned by second movers in each part and treatment. For the frequency of second movers sending 0, 50 or 150 points see Appendix B.

3.2. Punishment and Retaliation

As Figure 3A illustrates (see also Table B1), a large number of subjects are willing to spend some or all their monetary gains in order to either punish second movers or

[6] Throughout the paper, unless otherwise noted, we always use a two-sided Wilcoxon-Mann-Whitney test. We use each subject as an independent observation for tests concerning either part 1 or part 2. If we combine the data of both parts to perform a test, for each subject we first calculate the mean for the variable in question and then compute the test using these means as the independent observations. There are subjects from whom we have data from only one of the parts for various variables (e.g. a second mover who faces a first mover who cooperates in part 1 and a first mover who defects in part 2). In these cases, we take the data from the part for which we have information as that subject’s mean.

retaliate against first movers. In fact, around one third of the cases in which first and second movers interact wind up in punishment by the first movers. If returns were less than 150 points, about two thirds of the interactions end up in punishment (68.1%). When given the opportunity, retaliation by second movers is somewhat less frequent (40.0%). We even observe that, of the first movers who had the chance to punish second movers who retaliated, 55.6% decided to do so (we refer to this as 'additional punishment').[7]

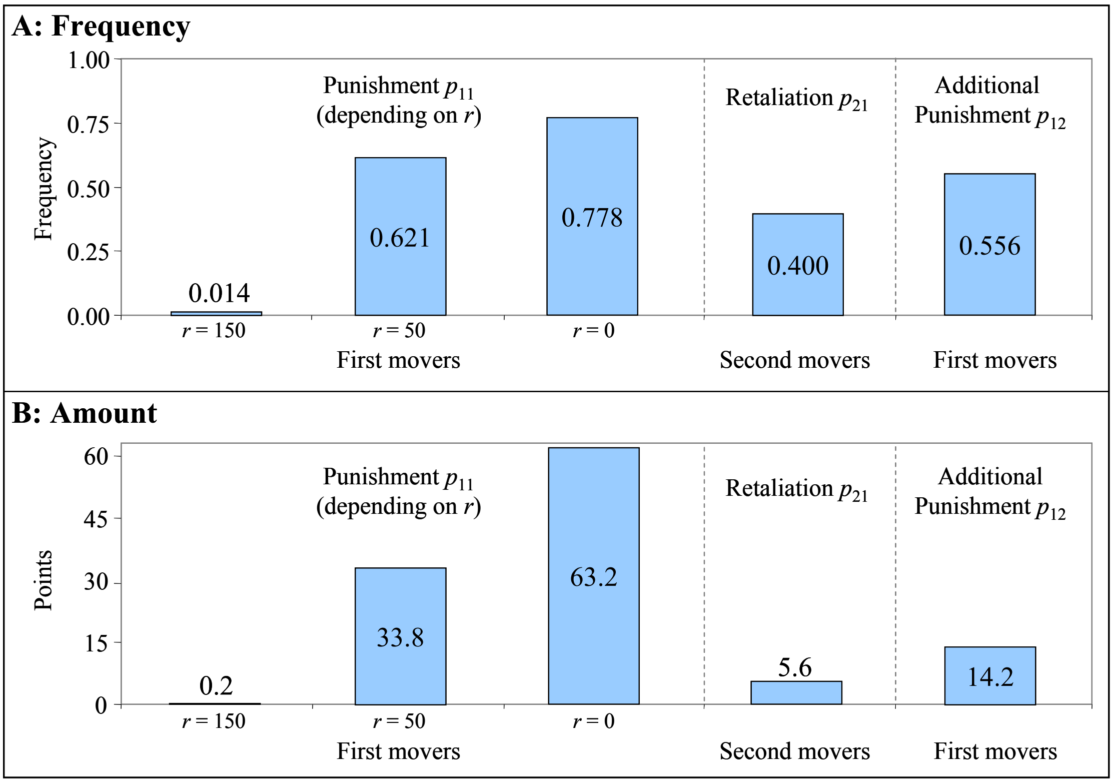

FIGURE 3 – PUNISHMENT AND RETALIATION

Note: A) Frequency of punishment (p_{11}), retaliation (p_{21}), and additional punishment (p_{12}) over both parts. B) Mean amount of points spent on punishment (p_{11}), retaliation (p_{21}), and additional punishment (p_{12}) over both parts.

Figure 3B shows that the amount spent on punishment by first movers who got back less than 150 points was clearly higher than the amount spent on retaliation by second movers who got punished ($p = 0.002$). Surely, since the earnings of first movers when they faced retaliation were lower than the earnings of second movers when they faced

[7] We only observe one case in which the second mover retaliated once again ($p_{22} > 0$). However, this is because in all the other pairs where the first mover punished a second time ($p_{12} > 0$) at least one of the players ended up with zero points and hence the punishment stage ended automatically.

punishment, this is partly explained by the ability of first movers to spend more on reducing the other's payoff. Still, if we normalize both punishment and retaliation using the maximum amount of points that an individual could assign to the other, we see that first movers are more aggressive punishers than second movers ($p = 0.080$).

Although it is not predicted by traditional economic theory (assuming own-payoff maximization), the punishment behavior of first movers is not surprising given that similar behavior has been observed in numerous experiments (see Camerer, 2003). Similarly it is consistent that the amount and frequency of punishment increases as the amount returned decreases.[8]

We find more unexpected the willingness of second movers to retaliate. After all, these subjects had behaved unkindly by returning less than 150 points. Furthermore, when they had to decide whether they wanted to retaliate, 65.0% of second movers had earnings that were actually higher or equal to the earnings of the first mover. It is remarkable that 7 (i.e. 53.8%) of these 13 second movers chose a positive amount of retaliation.[9] Unlike for first movers, we find that the retaliatory behavior of second movers does not depend on the actions of the other player. For instance, there is no significant difference in the amount or the frequency of retaliation between second movers who received a large amount of punishment and second movers who received a small amount (punishment above and below the median, $p > 0.355$).

It is instructive to calculate how retaliation affects the first movers' 'real' cost of punishment. Whenever first movers punish, they not only incur the cost of reducing the second mover's earnings, but they also risk further losses if the second mover decides to retaliate.[10] If there is no retaliation, the cost of punishment is 0.250 points per point reduced. Including the actual losses due to retaliation shows that, on

[8] Comparing first movers who received 150 points with first movers who received 50 or 0 points gives a significant difference for both the amount and the frequency of punishment (in each part $p < 0.001$). If we compare the amount and frequency of punishment of first movers who received 50 points with that of those who received 0 points, we find a significant difference only for the amount of punishment in the second part ($p = 0.020$, and in all other cases $p > 0.193$).

[9] This behavior is akin to 'misdirected' punishment in public good games. That is, punishment of high contributors by free-riders (Cinyabuguma et al., 2004; Gächter and Herrmann, 2005).

[10] The only case in which second movers cannot retaliate after being punished occurs when first movers who get back 0 points spend all of their remaining earnings punishing the second mover. In this case, both subjects end up with 0 points and no further retaliation is possible. Overall, 24.3% of the cases in which there was positive punishment fit this description.

average, first movers lost an additional 0.149 points per point reduced. This is a substantial increase of 59.4% in the cost of punishment. A similar analysis for the real cost of retaliation (given losses due to additional punishment) gives that second movers incur an additional 0.763 points per point reduced. This is a remarkable 305.6% increase in the cost of retaliation.[11] We now turn to how first and second movers adjust their behavior from part 1 to part 2.

3.3. Dynamics

As already noted, the starkest difference between treatments concerning the behavior of first movers is the large decrease in cooperation from part 1 to part 2 in the baseline treatment compared to the punishment treatment. On closer inspection, this difference is due to two reasons. First, as shown in Figure 4, first movers in the baseline treatment who got back less than 150 points in part 1 were more likely to defect in part 2 compared to first movers in the punishment treatment (p = 0.013). Second, in the baseline treatment more second movers chose to return less than 150 points (81.8% in the baseline treatment vs. 35.6% in the punishment treatment, p = 0.005). Hence, it appears that punishment has two desirable effects. On one hand, second movers anticipate punishment and as a result increase the amount returned. On the other hand, after experiencing opportunistic behavior, first movers are more willing to keep on cooperating if they have the opportunity to punish. In fact, if we examine how first movers in the punishment treatment adjust their behavior, we find that, among the first movers who received less than 150 points, those who actually punished are less likely to stop cooperating than those who did not punish (p = 0.087, see Figure 4).

[11] In fact, these calculations include pairs of subjects where no more punishment or retaliation was possible given that earnings were less than or equal to zero (e.g. see footnote 10). Excluding these observations raises the cost of punishment by 0.196, a 78.4% increase, and the costs of retaliation by 0.849, a 339.5% increase.

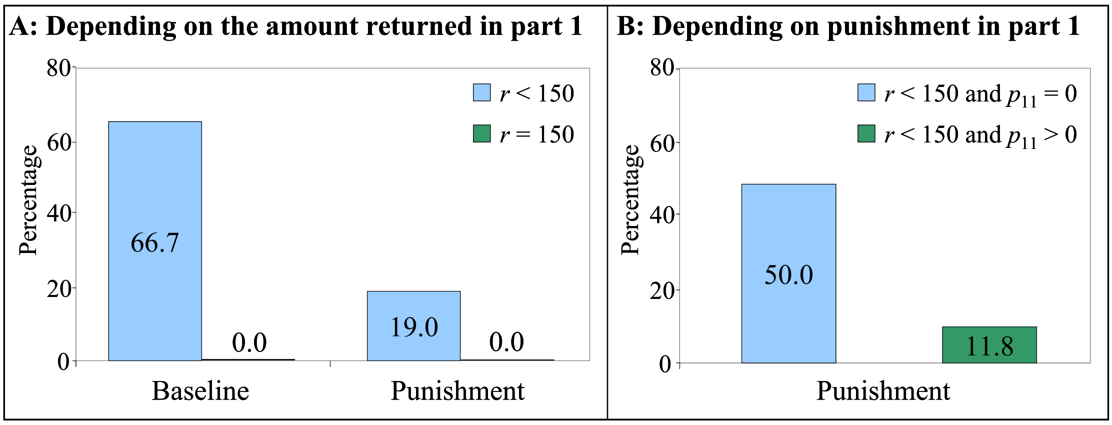

FIGURE 4 – DEFECTION IN PART 2 DEPENDING ON THE EVENTS IN PART 1

Note: A) Percentage of first movers who defect in part 2 depending on the amount returned by the second mover of part 1 in each treatment. B) Percentage of first movers who defect in part 2 depending on whether or not they punished the second mover of part 1 for returning less than 150 points.

We find a less clear pattern when we look at how second movers adjust their behavior. In both treatments, when given the opportunity, the majority of second movers choose the same action in both parts (80.0% in the baseline and 75.0% in the punishment treatment). Of those who change their decision, most of them decrease the amount returned (100.0% in the baseline and 84.6% in the punishment treatment). In order to look at the effect of punishment, we concentrate on second movers who had a good chance of being punished (i.e. those who returned less than 150). We find that, on average, second movers who were not punished decrease their returned amount by 25.0 points whereas those who were punished increase it by 10.0 points (p = 0.113). The main findings from the behavioral data are summarized in the following result:

RESULT 1 – *In the presence of punishment opportunities, cooperation is sustained at high levels. This is because, second movers return more and first movers who punish do not stop cooperating after experiencing opportunistic behavior. Punishment of opportunistic behavior is common despite the fact that in numerous cases punishment leads to various rounds of reducing each other's earnings.*

4. Emotions and Punishment

In this section, we first examine which of the first movers' emotions are related to punishment. We find that anger-like emotions explain why some first movers punish while others do not. Subsequently, we concentrate on anger and analyze what triggers first movers to feel high intensities of this emotion.

4.1. Anger and Punishment

Throughout the experiment, first movers experienced a variety of emotions. However, we find that anger-like emotions are the only ones that are clearly related to the punishment decision. First movers that felt high intensities of anger-like emotions punish more than those who felt low intensities of these emotions. Furthermore, we also find that differences in anger-like emotions can explain why, after receiving retaliation, some first movers punish again while others do not.

As is illustrated in Figure 5, first movers who, after observing the amount that was returned by the second mover, felt high intensities of anger punish more and more often than first movers who felt low intensities of anger ($p < 0.001$ for both part 1 and 2).[12] Similarly, on average, after observing the amount of retaliation assigned to them by the second mover, first movers who felt angry punish more and more often than first movers who did not feel as angry (the difference is significant for the amount of additional punishment $p = 0.096$, but not for its frequency $p = 0.322$).[13]

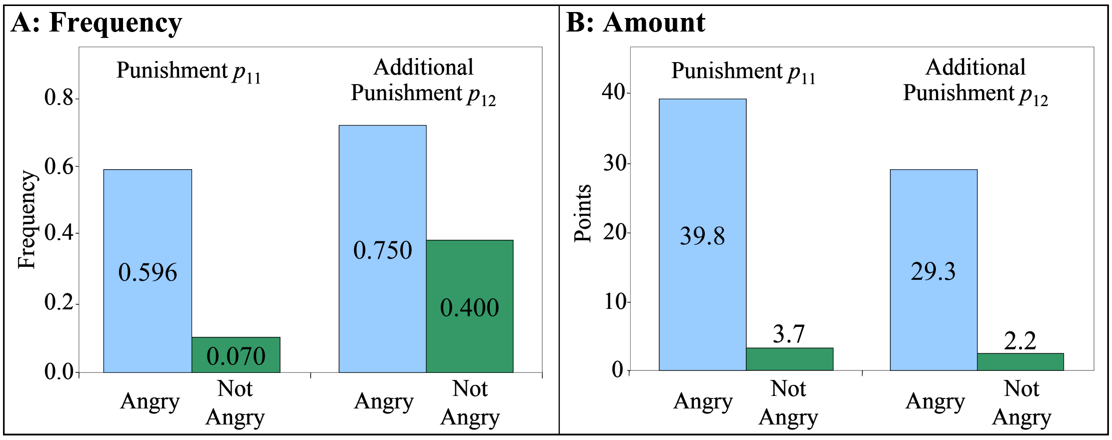

FIGURE 5 – ANGER AND PUNISHMENT

Note: A) Frequency of punishment by first movers depending on anger. B) Mean amount of points spent on punishment by first movers depending on anger.

[12] In the following analysis we will refer to a person feeling 'angry' if the reported value for anger was above the median, and as feeling 'not angry' if the value was below the median. Likewise for other emotions.

[13] Throughout this section, we report the results of tests done with the emotion of anger. However, we find very similar results and significance levels if we use irritation or (lack of) happiness.

Having found that punishment is related to experienced anger, the question arises what explains the different intensities of anger. We answer this question in the following subsection.

4.2. Causes of Anger

Anger experienced after observing the amount sent back by the second mover is caused by returns of less than 150 points, especially if they were unexpected or considered unfair (the emotional reaction of first movers to the amount returned can be found in Appendix B).

In both treatments, the most important trigger of high intensities of anger is simply receiving back less than 150 points. First movers in the punishment treatment who received 150 points felt lower intensities of anger than first movers who received either 50 or 0 points back ($p < 0.001$, see Table B3). Moreover, although on average first movers who received 0 points were angrier than those who received 50 points, the difference is significant only in the second part ($p = 0.328$ for part 1 and $p = 0.075$ for part 2).

In addition to the returned amount, the first movers' expectations have an effect on the intensity of anger. In particular, first movers who overestimated the amount returned by the second mover tended to be angrier than first movers who underestimated it. For example, if we control for the amount that was actually returned by concentrating on first movers who got back 50 points, we find that first movers who were expecting back 150 points were angrier than first movers who were expecting back 50 or 0 points (in each part $p < 0.043$).

Lastly, we also observe that fairness perceptions influence the amount of anger experienced by first movers. First movers who thought it is unfair to return low amounts were angrier than those who thought that it is fair to return low amounts (below or above median fairness). For instance, if we look again only at first movers who got back 50 points, we find that first movers who thought returning 50 was unfair were angrier than first movers who thought returning 50 was fair ($p = 0.004$).[14]

[14] We get similar results in a regression. Specifically, we estimate anger using the returned amount, the expected returned amount, and the perceived unfairness of returning 50 points as independent variables. We find a negative coefficient for the returned amount ($p = 0.001$) and positive coefficients for the other two variables ($p = 0.078$ and $p = 0.051$). This indicates that first movers feel angrier the less is returned, especially if they expected a high return or considered low returns to be unfair. Ordered probit estimates using robust standard errors and clustering on each subject, $\chi^2(3) = 74.4$.

Focusing on the emotional reaction of first movers to the amount of retaliation received from the second mover gives a comparable finding. Namely, first movers who faced no retaliation felt lower intensities of anger than first movers who faced positive retaliation (see Table B4, $p = 0.037$). Unfortunately, in this case we do not have enough observations to test for the effects of expectations and fairness perceptions. The findings of this section are summarized in the following result.

RESULT 2 – *First movers who punish do so because they are angry. High intensities of anger are triggered by opportunistic behavior by the second mover, especially if it is unexpected and considered unfair. Retaliation by second movers also makes first movers angry and leads to additional punishment*.

5. Social Emotions and Retaliation

We now turn to the relationship between the emotions and behavior of second movers. To begin with, we investigate the relationship between the emotions of second movers and their decision to retaliate. We also analyze whether emotions influence how second movers adjust their behavior over time. Next, we try to explain the difference in the emotional reactions of second movers.

5.1. Shame and Retaliation

As with first movers, the emotional reaction of second movers seems to be clearly related to their behavior (the emotional reaction of second movers can be found in Table B5). In particular, second movers who felt no shame are more likely to retaliate than other second movers. Furthermore, we also find that, for second movers who were punished, experiencing shame induces them to correct their behavior.

As can be seen in Figure 6A, second movers who felt no shame after being punished are more likely to retaliate than second movers who felt shame ($p = 0.045$).[15] We also get a similar result if we test for differences in the amount of points spent on retaliation ($p = 0.091$).

[15] We only report the results of tests using shame. However, for all findings in this section, we get very similar results and significance levels if we use guilt instead of shame.

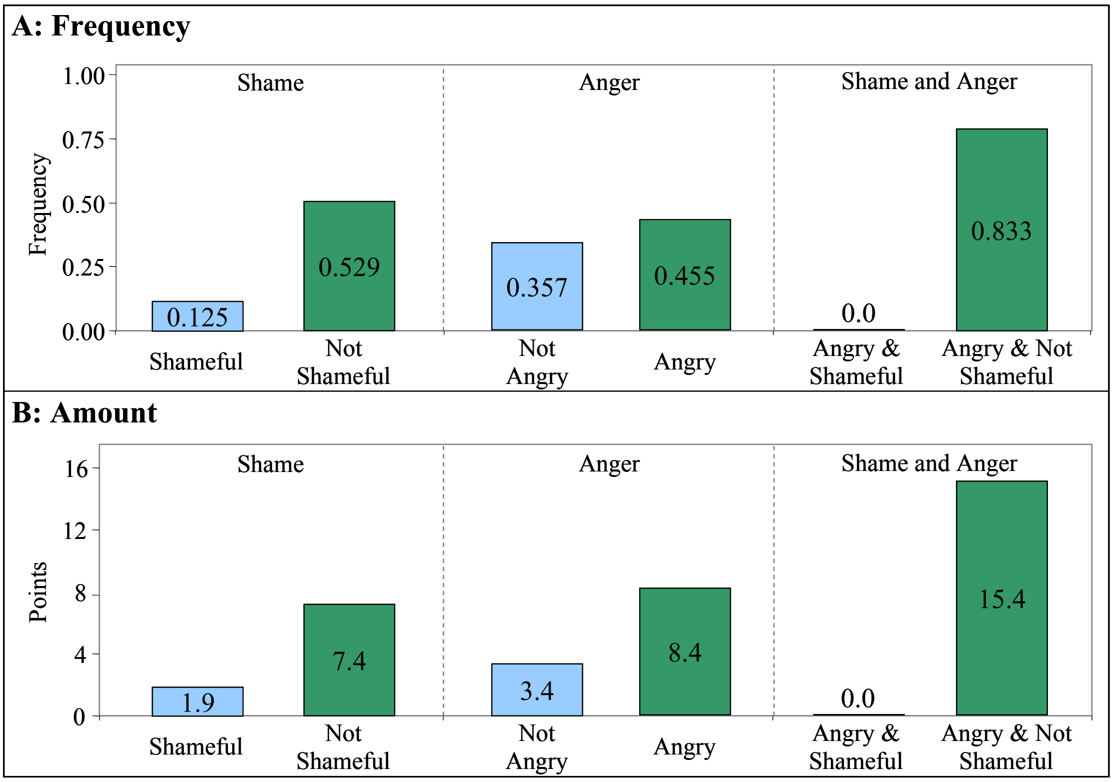

FIGURE 6 – SHAME, ANGER, AND RETALIATION

Note: A) Frequency of retaliation by second movers depending on anger and shame. B) Mean amount of points spent on retaliation by second movers depending on anger and shameful.

Interestingly, we also find that anger has an effect on the second movers' decision to retaliate. However, in this case the effect is not as straightforward. A simple look at the relationship between anger and retaliation, suggests that second movers who are angry retaliate more and more often than second movers who are not angry (see Figure 6). However, these differences are not significant (p = 0.739 when testing for differences in the amount of retaliation and p = 0.965 for differences in frequency). The effect of anger becomes obvious once we examine the interaction of anger and shame. In this case, a clear result is obtained. Namely, second movers who were angry and felt no shame retaliate more and more frequently than second movers who were angry and felt shame (p = 0.032 and p = 0.024). For second movers who were not angry, there are no significant differences between those who felt no shame and those who did (p > 0.637).

In addition to retaliation, shame is also related to how second movers adjust their behavior from part 1 to part 2. In Section 3.3 it was shown that second movers who were punished tend to return more in the subsequent part than second movers who were not punished. However, this difference is not significant. The emotional reaction of second movers reveals that the propensity of second movers to adjust their behavior after being punished depends on whether they felt shame or not. On average, second movers who felt shame after being punished increase the amount returned by 35.7 points whereas those who felt no shame decrease the amount returned by 12.5 points ($p = 0.053$). Since most second movers who returned less then 150 points were punished, we do not have enough observations to test the effects of shame on subjects that were not punished.

In conclusion, our results suggest that high intensities of anger provide second movers with a motivation to retaliate and high intensities of shame restrain them from doing so. Furthermore, shame seems to be necessary for punishment to have an effect on how second movers adjust their behavior. Next, we explain the differences in the intensities of anger and shame experienced by second movers.

5.2. Causes of anger and shame

The experience of anger among second movers depends on how many points they sent back to the first mover and on the amount of points the first mover spent punishing them. That is, second movers felt high intensities of anger if they received a high amount of punishment from the first mover. Furthermore, the intensity of anger is stronger the higher the amount they had returned before getting punished.

The most important reason why second movers get angry is simply receiving a positive amount of punishment (see Table B5). For example, second movers who were punished at least once reported significantly more anger than those who were never punished ($p = 0.001$).[16] Interestingly, if we examine whether the amount of punishment has an effect on anger we do not find a significant result. For example, second movers who were punished by a very large amount were not significantly angrier than those who were punished by a very small amount (top versus bottom quartile, $p = 0.624$). However, once we take into account the amount the second mover returned, we find a clearer effect. Among second movers who returned 50 points, those who were punished by a very large amount were angrier than those who were punished by a very small amount (top versus bottom quartiles, $p = 0.133$). The same pattern exists for second movers who returned 0 points (this time, $p = 0.168$).

[16] This is also true if we restrict ourselves to second movers who returned less than 150 points (in this case, $p = 0.002$).

For low amounts of punishment, second movers who returned 50 points were angrier than those who returned 0 points.[17]

More concisely, second movers became angry whenever they were punished, but if they had returned 50 instead of 0 points, they got angry at lower punishment amounts.[18] This is understandable given that second movers who returned 50 points not only behaved somewhat nicer than those who returned less, they also had lower earnings. Unlike first movers, we do not find that fairness perceptions or expectations (about the amount of punishment) have an effect on anger.

Unlike anger, it is not so clear what triggers different intensities of shame. We find that second movers who returned 150 points reported lower intensities of shame than those who returned less (in both treatments $p < 0.001$). In the punishment treatment, this is true even when we control for whether or not the second mover faced punishment. Specifically, second movers who returned 150 points and were not punished felt lower intensities of shame than second movers who returned less and were not punished ($p = 0.001$). In fact, punishment seems to have little effect on shame. For example, among second movers who returned less than 150, there is no significant difference in the amount of shame reported by those who were punished and those who were not ($p = 0.602$). Again, we do not find an effect of expectations or fairness perceptions on the experience of shame. The findings of this section are summarized in the following result.

[17] For instance, among second movers who did not receive very high punishment (i.e. excluding the top quartile), second movers who returned 50 points were more likely to feel angry (above the median) than those who returned 0 points ($p = 0.083$).

[18] These effects are more clearly captured in a regression. We estimate anger using the following independent variables: the amount returned, the expected amount of punishment, the perceived fairness of returning 50 points, and three variables I^r for $r \in \{0, 50, 150\}$ where $I^r = 0$ if the amount returned was different from r and I^r = amount of punishment if the amount returned was r. We obtain positive and significant coefficients for I^0, I^{50}, and I^{150} ($p < 0.001$). Furthermore, the coefficients are all significantly different from each other, with the coefficient for I^0 being the smallest and the one for I^{150} being the largest (Wald tests, $p < 0.009$). This suggests that for a given amount of punishment, second movers are angrier the more they had returned. Ordered probit estimates using robust standard errors and clustering on each subject, $\chi^2(6) = 73.8$.

RESULT 3 – *Second movers who retaliate do so because they are angry and do not feel shame. In addition, following the feeling of shame, second movers rectify their opportunistic behavior. High intensities of anger are triggered by punishment, especially if the second mover had returned a positive amount. High intensities of shame are triggered by opportunistic behavior and are not affected by punishment.*

6. Discussion and Conclusions

In this paper, we have shown that a realistic punishment institution, in which multiple rounds of punishment and retaliation are possible, is an effective tool for the support of cooperative behavior. However, retaliation is a commonly observed behavior that often results in the extreme reduction of the payoffs of the individuals involved. Furthermore, we have confirmed, that anger-like emotions are an important motivation for punishment. Opportunistic behavior induces anger and thus increases the likelihood of punishment. Lastly, we have shown that the experience of prosocial emotions, namely shame and guilt, restrain angry individuals from retaliating. Therefore, prosocial emotions can be seen as a mechanism managing the behavioral reactions of anger.

Given that costly punishment has been shown to be an effective way of enforcing cooperative behavior, it is important to have a good understanding of the motivations and reactions of both the punishers and the punished. We find interesting that individuals who are willing to punish are also willing to keep on cooperating (see Result 1). This guaranties that, as long as these individuals have the opportunity to punish, cooperation can be sustained. Furthermore this kind of individuals might help cooperation emerge, even if it was initially rare. In addition, the same type of people is necessary to support punishment in the presence of retaliation. If retaliation deters individuals from using the punishment mechanism, cooperation can unravel (Nikiforakis, 2004). However, if the opportunity to punish back always exists, this could prevent retaliation from limiting the punishment of opportunistic behavior.[19]

[19] Unfortunately, we do not have enough observations to determine if retaliation deters punishment. Only two first movers experienced both retaliation in part 1 and a second mover who returned less than 150 points in part 2. Given that both of these individuals punished the second mover in part 2, it appears that retaliation did not have much of an effect on them.

As expected, we find that the main motivation for the punishment of opportunistic behavior is experiencing anger. Furthermore, we confirm that individuals feel angrier the more money the other player took (Bosman and van Winden, 2002), the more unexpected was the opportunistic behavior (Ben-Shakhar et al., 2004), and the more strongly the individual felt about fairness (Pillutla and Murnighan, 1996). In fact, our results show that each of these motivations has a separate effect on the intensity of anger and thus on the propensity to punish.

Knowing that punishment is triggered by the emotion of anger can help us model this type of behavior. Since the action tendency of anger is to attack (Lazarus, 1991), and thus to harm whoever is negatively affecting our interests, punishment can be seen as the consumption of a good from which pleasure is derived (Quervain et al., 2004). Interpreting punishment as simply another good allows us to apply standard theoretical economic analysis to an otherwise puzzling phenomenon (see Carpenter, 2004). It is important to point out that, even if anger was triggered by unfair behavior (e.g. deviations from equality or a maximin norm), the goal of angry individuals is to harm the other party, and not, through punishment, to correct unfair material outcomes.[20] For example, if in our game first movers who got back 50 points used punishment to rectify an unfair distribution of income, they should not spend more than 66.67 points on punishment (this amount gives both players equal earnings). However, a substantial number of first movers punish more than this amount.[21] In this sense, outcome based models of social preferences such as Fehr and Schmidt (1999), and Bolton and Ockenfels (2000) miss an important characteristic of punishment behavior (see also Reuben and van Winden, 2004).

An important and yet overlooked aspect of punishment is the emotional reaction of the punished. As was shown in this paper, prosocial emotions such as shame play a crucial role for the viability of punishment for the enforcement of social norms. In Section 5 we have shown that feeling shameful helps explain why some individuals who acted selfishly adjust their behavior whereas others do not. It has been observed that in public good games, the use of non-monetary punishment has a

[20] In this respect, as is argued by Carpenter and Matthews (2005), there is an important difference between anger-induced punishment by the affected individual and indignation-induced punishment by an unaffected third party.

[21] To be precise, 31.3% of the first movers who punished after receiving 50 points back punished, at least once, by more than 66.67 points.

positive effect on contribution levels.[22] However, our results indicate that, it is the combination of feeling shame and receiving monetary punishment that has a significant effect on behavior. This suggests that shame alone will not have an effect if the cooperative norm is not actively enforced. Hence, although non-monetary punishment has the desirable property that it can affect behavior without destroying resources, the lack of real consequences for free-riders might make this effect deteriorate over time (Masclet et al., 2003). In this sense, as is shown by Noussair and Tucker (2005), the best performing punishment institution is one in which both symbolic and monetary punishments are available.

Another essential role for shame is the prevention of retaliation by punished individuals. As was shown in Section 4, even if they acted unkindly, individuals do feel angry when they are punished. However, it is only those individuals who are angry and do not feel shame that decide to retaliate. Therefore, if it were not for some individuals experiencing shame, retaliation would be much more common and punishment of selfish behavior much more costly. For example, if second movers who felt shame had behaved as second movers who felt no shame (controlling for anger) then retaliation would have been 42.6% more frequent and 50.6% higher. Furthermore, the decrease in the amount returned from part 1 to part 2 would have been 48.8% bigger. Social emotions like shame are thus essential for the effectiveness of a punishment institution. This supports the assumption that social emotions coevolved with institutions and anger-like emotions in order to limit antisocial actions (Bowles and Gintis, 2001). An interesting question for further exploration is the specific evolutionary mechanisms that lead to this situation.

Finally, even though we did not differentiate in our analysis between shame and guilt, we would like to stress that the action tendencies of the two emotions can be different (Tangney and Dearing, 2002). On one hand, shame is related to a devaluation of the self, and therefore the action tendency of shame is withdrawal and avoidance of further contact. On the other hand, guilt is more related to the blameworthiness of an act and is thus more likely to result in reparation and action.[23]

[22] For instance, Masclet et al. (2003) use symbolic punishment points and find that, in the short run, they work almost as well as real punishment points. Barr (2001) reports that the public blaming of the free-rider can increase cooperation in future rounds.

[23] Economists usually distinguish shame and guilt by the visibility of behavior. Shame is said to be triggered in social situations in which actions are seen by others, whereas guilt is more related to internalized values and hence is not influenced by the presence of others (e.g. Kandel and Lazear,

Therefore, if an outside option is available in which the experience of shame can be avoided, anticipation of feeling shame might have undesired effects on the prevalence of prosocial behavior (Lazear et al., 2005). In other words, when trying to decrease the frequency of selfish behavior, the attempt to explicitly induce shame, might result in avoidance of further interaction instead of in more cooperation.

Appendix A

These are the instructions for the first movers used in the punishment treatment. The instructions for the second mover and for the baseline treatment are available upon request.

A.1. Instructions

Part 1

There are two types of participants in this part, participants A and participants B. Half of the persons participating in the experiment will be in the role of participant A, and the other half in that of participant B. *You are a participant A*.

In part 1 of the experiment, you will be randomly assigned a participant B. During this part, you will interact only with this participant B. Moreover, you will *not* interact again with this participant in part 2 of the experiment. Part 1 consists of three steps. In step one, you must decide whether you will transfer points to participant B or if you will retain the points for yourself. In step two, participant B will decide if he will transfer points to you or if he will keep them himself. In step three, both of you must again make a decision. There are various options in step three, which will be explained below. We will also describe the exact experimental procedure on the next pages.

Procedure for the three steps

At the beginning of part 1 you and participant B will each receive 100 points as earnings.

Step one

1992). However, research by psychologists has shown that people feel shame even when their actions are unobserved (Tangney et al., 1996), and that the experience of guilt varies considerably depending on the interpersonal context (Baumeister et al., 1994).

At the beginning of the first step you will receive 50 decision points. Participant B will receive no decision points. In step one, you must decide whether you want to transfer your 50 decision points to participant B or transfer no points to participant B. If you transfer the 50 points, they will be multiplied by six, meaning that participant B will receive 6×50 = 300 points. Then, step two begins. If you decide to transfer nothing part 1 will end here.

Step two

In step two, participant B has to decide whether he will transfer 150, 50 or 0 points to you. You will then receive exactly the number of points B transferred.

Therefore, four possibilities exist after the first two steps:

	Your additional earnings	**B's additional earnings**
You retain your decision points.	50 points	0 points
You transfer your decision points and B transfers 150 points.	150 points	150 points
You transfer your decision points and B transfers 50 points.	50 points	250 points
You transfer your decision points and B transfers nothing.	0 points	300 points

Hence, after step two your total earnings will be:

100 + the additional earnings from the table above.

Step three

In step three, you will be informed how many points participant B transferred to you. Now, you can assign penalty points to participant B. The assignment of penalty points has financial consequences for both participants, A and B. Each penalty point which you assign costs you one point, while four points are deducted from your participant B. If you assign three penalty points to participant B, this will cost you three points and participant B will have twelve points deducted.

You cannot deduct more points from participant B than his total earnings in that part (i.e. 100 + B's additional earnings). If participant B has 250 points after step 2, then with your assignment of penalty points you can reduce his earnings by at most

250 points. Hence, as long as your participant B has positive earnings, you can assign him as many penalty points as you want. You can also assign him no penalty points.

Participant B will then be informed how many penalty points you assigned him and how many points were deducted from his earnings. If you decided not to assign penalty points, part 1 will end here. If you assigned penalty points to participant B, he can decide to assign penalty points to you. The assignment of penalty points has the same financial consequences as described above. Each penalty point that participant B assigns to you costs him one point, while four points are deducted from your earnings. You can not be deducted more points than the total earnings you own at that moment. If participant B decides to assign no penalty points to you, part 1 will end here. Note: Participant B can assign penalty points even if his earnings at that point are zero. If he does so, he will lose points in part 1 of the experiment.

If participant B assigned you penalty points, you and participant B will have the option to assign penalty points to each other in turns. Part 1 will end when either you or participant B decides to assign no penalty points, or if either you or participant B can not be assigned penalty points because your or his earnings are zero or less. In other words, as long as one of you assigns a positive amount of penalty points, the other will have the opportunity to assign penalty points back. Note that, you will be able to assign penalty points *even if your earnings at that point are zero*. Furthermore, you *cannot* be assigned penalty points if your *own* earnings are zero.

Finally

Remember that, you participate in part 1 only *once*. Therefore consider your decisions carefully. At the end of part 1 you will receive instructions for part 2 of the experiment.

Part 2

We will now give you the instructions for part 2 of the experiment.

Also in this part there will be two types of participants, participants A and participants B. Every person participating in the experiment will be in the role they had in part 1. Therefore, *you are a participant A*. As in part 1 you will be randomly assigned a participant B. During this part, you will interact only with this participant B. You can be certain that *this participant B is not the same person as in part 1*.

This part will consist of the same three steps as part 1. Therefore exactly the same instructions apply for part 2 as for part 1. Remember that you will participate in this part only *once*. Therefore consider your decisions carefully.

Examples of questions in the self-reports

To measure emotions:

- Indicate how intensely you feel each of the following emotions right now, *after knowing the amount that B transferred to you*?

The subject then filled in a series of seven-point scales that ranged from ‘not at all’ (1) to ‘very intensely' (7).

To measure expectations:

- Player A can now assign you penalty points. How many penalty points do you think A will assign to you?

The subject then entered a point estimate.

To measure fairness perceptions:

- Suppose that participant A transfers the 50 decision points to participant B. Participant B has to choose to transfer back either 150 points, 50 points or 0 points. In your opinion, how *fair* do you believe is each of these choices: If participant B transfers back 150 (50, 0) points this choice is ... ?

The subject then filled in three seven-point scales (one for each choice) that ranged from ‘very unfair’ (1) to ‘very fair’ (7).

Appendix B

Table B1 and Table B2 summarize of the behavioral data for each treatment.

TABLE B1 – SUMMARY OF THE BEHAVIORAL DATA IN THE PUNISHMENT TREATMENT

Means	**Part 1**	**Part 2**	**Both parts**[24]
Points sent (cooperation)	43.4	43.4	43.4
standard deviation	(17.1)	(17.1)	(14.7)
Frequency of cooperation	86.4	86.4	86.4
Number of observations	68	68	68
Points returned	108.5	96.6	103.4
standard deviation	(58.1)	(62.9)	(57.5)
Frequency of returning 150	0.644	0.559	0.614
Frequency of returning 50	0.237	0.254	0.227
Frequency of returning 0	0.119	0.186	0.159
Number of observations	59	59	66
Points spent on punishment	17.3	18.7	18.1
standard deviation	(31.4)	(35.5)	(26.2)
Frequency of punishment	0.305	0.254	0.278
Number of observations	59	59	63
Points spent on retaliation	5.5	5.9	5.2
standard deviation	(8.7)	(10.0)	(8.2)
Frequency of retaliation	0.375	0.444	0.400
Number of observations	16	9	20
Points spent on additional punishment	6.2	24.3	14.2
standard deviation	(8.8)	(28.0)	(20.6)
Frequency of additional punishment	0.600	0.500	0.556
Number of observations	5	4	9

[24] To be precise the data in this column is the mean behavior of each subject across both parts. In other words, first we take the mean behavior across parts for each subject and then we take the mean across all subjects. In the cases where a subject had only one opportunity to take an action, we take the data from that part as that subject's mean.

TABLE B2 – SUMMARY OF THE BEHAVIORAL DATA IN THE BASELINE TREATMENT

Mean	Part 1	Part 2	Both parts[24]
Points sent (cooperation) standard deviation	42.3 (18.8)	23.1 (25.9)	32.7 (15.8)
Frequency of cooperation	84.6	46.2	65.4
Number of observations	13	13	13
Points returned standard deviation	36.4 (59.5)	41.7 (58.5)	35.4 (56.9)
Frequency of returning 150	0.182	0.167	0.167
Frequency of returning 50	0.182	0.333	0.208
Frequency of returning 0	0.636	0.500	0.625
Number of observations	11	6	12

The emotional reaction of first movers in the punishment treatment is summarized in Table B3 and Table B4. In the baseline treatment, the emotional reaction of first movers was statistically indistinguishable from the one in the punishment treatment. It seems that the opportunity to punish does not affect how first movers feel about the amount returned to them by second movers.

TABLE B3 – MEAN EMOTIONAL INTENSITY OF FIRST MOVERS AFTER OBSERVING THE AMOUNT RETURNED BY THE SECOND MOVER IN THE PUNISHMENT TREATMENT

Emotions	Got back 150	Got back 50	Got back 0
Anger standard deviation	1.1 (0.5)	4.5 (1.9)	5.8 (1.5)
Irritation standard deviation	1.2 (0.7)	5.0 (1.5)	6.1 (1.5)
Happiness standard deviation	6.1 (1.0)	2.3 (1.4)	1.8 (1.1)
Gratitude standard deviation	4.9 (1.8)	2.4 (1.7)	1.6 (1.1)
Shame standard deviation	1.2 (0.5)	1.9 (1.6)	2.9 (2.3)
Guilt standard deviation	1.1 (0.5)	1.3 (0.9)	1.8 (1.7)
Surprise standard deviation	4.2 (1.6)	3.9 (1.7)	4.5 (2.5)
Number of observations	53	27	17

TABLE B4 – MEAN EMOTIONAL INTENSITY OF FIRST MOVERS AFTER OBSERVING THE AMOUNT OF RETALIATION THEY RECEIVED FROM THE SECOND MOVER

Emotions	No Retaliation	Positive Retaliation
Anger	1.9	3.6
standard deviation	(1.5)	(2.2)
Irritation	2.2	4.7
standard deviation	(1.7)	(2.2)
Happiness	3.4	2.6
standard deviation	(1.8)	(1.3)
Gratitude	2.4	2.7
standard deviation	(2.0)	(1.9)
Shame	2.1	1.3
standard deviation	(1.8)	(0.9)
Guilt	2.1	1.5
standard deviation	(1.9)	(1.1)
Surprise	4.8	2.3
standard deviation	(1.9)	(1.6)
Number of observations	14	10

The emotional reaction of second movers is summarized in Table B5.

TABLE B5 – MEAN EMOTIONAL INTENSITY OF SECOND MOVERS AFTER OBSERVING THE AMOUNT OF PUNISHMENT THEY RECIVED FROM THE FIRST MOVER

Emotions	Not Punished	Below Median Punishment	Above Median Punishment
Anger	1.1	3.6	3.9
standard deviation	(0.8)	(2.2)	(1.9)
Irritation	1.3	3.5	4.8
standard deviation	(1.2)	(2.3)	(2.3)
Happiness	5.0	2.4	1.5
standard deviation	(1.6)	(1.4)	(0.8)
Gratitude	4.0	2.5	2.3
standard deviation	(2.0)	(1.5)	(1.7)
Shame	1.2	1.3	1.7
standard deviation	(0.9)	(0.6)	(1.1)
Guilt	1.4	1.8	1.9
standard deviation	(1.1)	(1.3)	(1.3)
Surprise	2.5	4.0	5.2
standard deviation	(1.9)	(2.1)	(2.1)
Number of observations	55	14	13

References

Anderson, S., A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature neuroscience*, 2: 1032-1037.

Barr, A. (2001). Social Dilemmas and Shame-based Sanctions: Experimental results from rural Zimbabwe. Working paper.

Baumeister, R. F., A. M. Stillwell, and T. F. Heatherton (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, 115: 243-267.

Ben-Shakhar, G., G. Bornstein, A. Hopfensitz, and F. van Winden (2004). Reciprocity and emotions: Arousal, self-reports, and expectations. Discussion paper 04-099/1. Tinbergen Institute.

Bolton, G. and A. Ockenfels (2000). A theory of equity, reciprocity, and competition. *American Economic Review*, 90: 166-193.

Bosman, R. and F. van Winden (2002). Emotional Hazard in a Power to Take Experiment. *The Economic Journal*, 112: 147-169.

Bowles, S. and H. Gintis (2001). The economics of shame and punishment. Working paper.

Camerer, C. (2003). *Behavioral Game Theory*. New Jersey: Princeton University Press.

Carpenter, J. P. (2004). The Demand for Punishment. Working paper. Middlebury College.

Carpenter, J. P. and P. Matthews (2005). Norm Enforcement: Anger, Indignation or Reciprocity. Working paper. Middlebury College.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics* 117: 817-869.

Cinyabuguma, M., T. Page, and L. Putterman (2004). On perverse and second-order punishment in public goods experiments with decentralized sanctioning. Working paper. Brown University.

Damasio, A. (1994). *Descartes' Error - Emotion, Reason and the Human Brain*. Harper Collins.

Dufwenberg, M. and G. Kirchsteiger (2005). A theory of sequential reciprocity. *Games and Economic Behavior*, forthcoming.

Egas, M. and A. Riedl (2005). The economics of altruistic punishment and the demise of cooperation. Working paper. University of Amsterdam.

Elster, J. (1999). *Strong Feelings: Emotion, Addiction and Human Behavior*. MIT Press.

Falk, A. and U. Fischbacher (2000). A theory of reciprocity. Working paper No. 6. Institute for Empirical Research in Economics, University of Zürich.

Fehr, E. and S. Gächter (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, 90: 980-994.

Fehr, E. and B. Rockenbach (2003). The detrimental effects of sanctions on human altruism. *Nature*, 422: 137-140.

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

Fischbacher, U. (1999). Zurich toolbox for readymade economic experiments, experimenter's manual. Working Paper No. 21. Institute for Empirical Research in Economics, University of Zurich.

Gächter, S. and B. Herrmann (2005). Norms of cooperation among urban and rural dwellers: Experimental evidence from Russia. Working paper. University of Nottingham.

Kandel, E. and E. P. Lazear (1992). Peer pressure and partnerships. *Journal of Political Economy*, 100: 801-817.

Lazarus, R. (1991). *Emotion and Adaptation*. Oxford University Press.

Lazear, E. P., U. Malmendier, and R. A. Weber (2005). Sorting in experiments. Working paper. Stanford University.

Loewenstein, G. (1996). Out of control: Visceral influence on behavior. *Organizational Behavior and Human Decision Processes*, 65: 272-292.

Masclet, D., C. Noussair, S. Tucker, and M. C. Villeval (2003). Monetary and non-monetary punishment in the voluntary contribution mechanism. *The American Economic Review*, 93: 366-380.

Moll, J., R. de Oliveira-Souza, P. J. Eslinger, I. E. Bramati, J. Mourao-Miranda, P. A. Andreiuolo, and L. Pessoa (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *The Journal of Neuroscience*, 22: 2730-2736.

Nikiforakis, N. S. (2004). Punishment and Counter-punishment in Public Good Games. Working paper. Royal Holloway University of London.

Noussair, C. and S. Tucker (2005). Combining Monetary and Social Sanctions to Promote Cooperation. *Economic Inquiry*, forthcoming.

Ortony, A., G. Clore, and A. Collins (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action: Presidential address, American Political Science Association, 1997. *American Political Science Review*, 92: 1-22.

Pillutla, M. and J. K. Murnighan (1996). Unfairness, anger and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes*, 68: 208-224.

Quervain, D. J. F., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr (2004). The neural basis of altruistic punishment. *Science*, 305: 1254-1258.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281-1302.

Reuben, E. and F. van Winden (2004). Reciprocity and emotions when reciprocators know each other. Discussion paper 04-098/1. Tinbergen Institute.

Robinson, M. and G. Clore (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128: 934-960.

Sanfey, A. G., J. K. Rilling, J. A. Aronson, L. E. Nystrom, and J. D. Cohen (2003). The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science*, 300: 1755-1758.

Tangney, J. P. and R. L. Dearing (2002). *Shame and Guilt*. The Guilford Press.

Tangney, J. P., R. S. Miller, L. Flicker, and D. H. Barlow (1996). Are shame, guilt and embarrassment distinct emotions? *Journal of Personality and Social Psychology*, 70: 1256-1269.

Thaler, R. (2000). From homo economicus to homo sapiens. *Journal of Economic Perspectives*, 14: 133-141.