

Costly Punishment as a Decision-Making Problem: Evidence from Economic Experiments

Simon Jonas Hadlich | Student no.: 6317170 | Email: simon.hadlich@student.auc.nl

Abstract

Prosocial punishment is a pervasive feature of human interactions. However, standard economic theory denies the possibility of other-regarding or irrational behavior. This paper reviews experimental studies using public goods games and similar settings with a punishment option. Prosocial punishment is found to occur across a large variety of populations and experimental settings. Furthermore, antisocial punishment is observed to be common in some population. Neurological studies show that cognitive, emotional, and motivational processes are involved in making decisions about costly prosocial punishment. This requires new models of decision-making in cognitive psychology.

Keywords: Behavioral Economics; Punishment; Altruism; Expected Utility Theory; Neuroeconomics

Introduction

Mainstream economic theory, as it is taught in colleges around the world, is based on expected utility theory - the assumption that human being will behave in a rational and self-interested way. While this world view has long been criticised, in particular by cultural and critical theorists, in recent years adversity comes increasingly from within economics.

Behavioural economists are using simple economic games to test the mathematical predictions of game theory - and thus expected utility theory - in experimental settings. By now, a plethora of studies have shown that game theoretical predictions do not hold true in interaction between humans. This also has important

implications for theories about the evolution of cooperation.

In particular, humans around the world have been shown to engage in costly punishment of defectors, even if they themselves have not been harmed. The possibility of punishment has been shown to effectively increase cooperation, but mainstream economic theory nevertheless predicts that it should not take place. This paper thus investigates the factors that guide decision-making about punishment in economic games.

Expected Utility Theory

Expected utility theory understands human decision-making as guided by rational self-interest. When presented with all relevant information, people will choose the option that has the maximum expected utility (Goldstein, 2011). This approach to decision-making can be cast in terms of game theory, where human choice will coincide with Nash equilibrium outcomes.

Economic Games

Scholars from various disciplines, including behavioural economists, evolutionary biologists, and economic anthropologists, have tested the assumptions of rationality and self-interested preferences made by the canon of economic theory in experimental settings. Several game settings have been used in these experiments, most notably the dictator game - often used as a measure of intrinsic altruism - and the ultimatum game. A more complicated game modelling economic relationships is the public goods (PG) game.

Some economic games, such as the ultimatum game, allow for punishment in the form of discontinued cooperation. Other games have been amended to include an explicit punishment opportunity following the initial transaction. In particular, Fehr and Gächter (2000) have developed a variation of the public goods game that allows players to punish individual group members after each interaction. A complex variation of the public goods game is the constituent common pool resource (CPR) game developed by Ostrom, Walker and Gardner (1992). A two-player game used in

neuroimaging studies is the social dilemma (SD) game with punishment condition (Fehr, Fischbacher, & Kosfeld, 2005). Another game, the third-party punishment (3PP) game used by Henrich et al. (2006), enables a third party not involved in the initial interaction to punish an individual player. Since the PG game is by far the most repeated of these three, this review will focus on this game.

The PG game was designed by Fehr and Gächter (2000) such that if the assumptions of rationality and selfishness hold, cooperation or punishment can never be part of a subgame-perfect equilibrium. In the PG game, each member of a group - most commonly of size four - receives some initial endowment of money units. They can then decide whether to keep these or to invest them (partly or fully) in a common project. All contributions are multiplied by some factor and then split up equally among the players. Under this condition, free-riding is a dominant strategy, although aggregate pay-off is maximized if all players fully cooperate. The punishment condition adds a second stage to this game, at which subjects are informed about the individual contributions. They then have the opportunity to punish individual group members at a cost to themselves. Again, if the assumptions of rationality and self-interestedness hold, no player would pay to punish another player, and thus free-riding would be the dominant strategy under the punishment condition.

Both conditions of the PG game, with and without punishment, have frequently been compared to understand the impact of punishment on cooperation. Apart from non-punishment or punishment, several further conditions can be altered to adjust the PG game. Most studies use repeated iterations of PG games; Gächter, Renner, and Sefton (2008) studied the differences between short-run (10 iterations) and long-run (50 iterations) PG games. Fehr and Gächter (2000) used both changing ("stranger" treatment) and constant ("partner" treatment) groups; the former is essentially a series of one-shot games. While Fehr and Gächter implement total anonymity, other studies allow for reputation-building (Nowak & Sigmund, 2005). de Quervain et al. (2004) compared effective and symbolic punishment, where the latter is not costly to the punished player. Furthermore, the number of players, the multiplier for the investment in the public good, and the cost of punishment differ across studies.

Punishment in Economic Games

Punishment Behaviour

One of the most persistent findings in experimental economics has been that people are willing to spend considerable amounts of money to punish other players, even if this is not in their rational self-interest. Evidence for costly punishment exists from various cultures (Herrmann, Thöni, & Gächter, 2008), and even holds in third-party punishment games, where the punisher is not part of the initial interaction (Henrich et al., 2006).

Fehr and Gächter (2000) find that punishment in PG games is linked to individual contributions' relation to the group average. In both stranger and partner treatments, players were more severely punished the more their contribution fell below the group's average contribution. This behaviour has been termed "pro-social punishment", as it targets players which benefit from free-riding on other players' contributions. The prevalence of pro-social punishment has repeatedly been replicated (e.g. Fehr & Gächter, 2002), and pro-social punishment behaviour in PG games does only weakly significantly differ across societies (Herrmann, Thöni, & Gächter, 2008).

In contrast, Fehr and Gächter (2000) did not find systematic punishment of contributions above the average, i.e. antisocial punishment. Cross-cultural studies, however, provide evidence that the absence of antisocial punishment is not universal. Herrmann, Thöni, and Gächter (2008) studied punishment behaviour in PG games across 15 locations. While antisocial punishment was all but absent in some participant pools, in other pools people punished players who contributed as much or more than them just as harshly as free riders.

People even punish others at a cost to themselves if they did not have a stake in the initial interaction. This is a finding from 3PP games administered by Henrich et al. (2006). They found considerable differences in punishment behaviour across 16 small-scale populations. Pro-social punishment is common in most of the participant pools, and generally decreased as offers approached equal sharing. However, the frequency of punishment differed widely. In some pools, less than a third of the participants punished even the most severe free riders, while in other pools, full free-riding was punished every time; on average two-thirds of players were willing to punish most

severe free-riding at a cost to themselves. In this study, antisocial punishment (defined as punishment for sharing more than equally) was rare, and occurred only in a few populations.

There is some evidence that the severity of punishment differs little depending on whether the punisher was part of the initial exchange or not. Strobel et al. (2011) compared punishment behaviour in first-person and third-party punishment conditions, and found only weakly significant differences between both treatments (caused by deviations under one condition). I.e. people punishing a player who free-rode on another, unrelated player would punish just as harshly as people directly cheated on by the free-rider themselves. However, this comparison has not been made for further and in particular non-Western populations.

Numerous studies have shown that many people are willing to punish others at a cost to themselves; a finding that holds true across diverse populations and various games, including first-person and third-party punishment conditions. Nevertheless, there is considerable heterogeneity in punishment behaviour within populations: behaviour which is punished by some people will be punished by others, even within the same population.

Punishment and Cooperation

A frequent finding in studies on public goods games is that the punishment condition increases cooperation. In simple, non-punishment public goods games, contributions decline over time, reaching very low levels towards the end of the game even if they start out high (e.g. Fehr & Schmidt, 1999). When punishment is possible, however, cooperation can be maintained, and oftentimes even increases. This has been found independent of whether groups are stable or changing (Fehr & Gächter, 2000). In Fehr and Gächter's early study, the "partner treatment" with punishment led to a convergence towards full cooperation (ibid.). The same study also found significantly higher initial contributions in the first round, meaning that the threat of punishment has an immediate effect on behaviour

In some studies, while punishment increased cooperation as compared to a non-punishment condition, it lowered overall pay-offs. Gächter, Renner, and Sefton (2008) show that this effect vanishes in the long run. Comparing PG games over 10 and 50

rounds, they found that the punishment condition increased cooperation in both cases; however, average net earnings were lowered by punishment in the short run. Under the long run condition, average net earnings were higher, in part because the prospect of many iterations already increased cooperation in early periods. Hence, the authors conclude that "punishment not only increases cooperation, it also makes groups and individuals better off in the long run."

While the cooperation-enhancing effect of punishment has been shown in a large number of experiments, cross-cultural studies indicate that it is dependent on environmental variables. A large-scale study by Herrmann, Thöni, and Gächter (2008) sampled subjects from 15 locations around the world. While Fehr and Gächter's (2000) previous findings were confirmed for some locations, punishment did not have a positive effect on cooperation in a number of other locations. This was linked to a high prevalence of antisocial punishment. Herrmann, Thöni, and Gächter (2008) show that antisocial punishment is associated to weak norms of civic cooperation and rule of law. In addition, Gächter, Herrmann, and Thöni (2010) have explored the link between culture and cooperation, finding that punishment did not increase cooperation in Southern European and Arabic-speaking locations.

Explaining Costly Punishment

As costly punishment strongly violates the assumption of rational self-interestedness, alternative explanations for the above observations are necessary. Fehr and Gächter (2000) posited that "subjects strongly dislike being the 'sucker'," i.e. to let others free-ride on their cooperative behaviour. Consequently, people would punish others who defected on their cooperation. However, Henrich et al.'s (2006) cross-cultural study on 3PP games seems to refute this hypothesis. In all 15 societies studied, at least some share of participants was willing to expend part of their endowment to punish players who free-rode on a third party.

The pervasiveness of pro-social punishment under both first-person and third-party punishment conditions points towards social preferences. Fehr, Fischbacher, and Kosfeld (2005) argue that when people engage in pro-social punishment, they exhibit a preference for equality (or inequality aversion). In this case, pro-social punishment

would be a rational course of action for individuals with non-selfish preferences.

For antisocial punishment, Fehr and Gächter (2000) provide five possible explanations. Apart from random error, they speculate that a) those who contribute above average might find others' contributions to low, even if they are above average, b) subjects may want to achieve a relative advantage over others, c) free-riders engage in spiteful revenge against cooperators as they expect to get punished by them, and d) previously punished subjects engage in blind revenge, assuming that they have been punished by cooperators and hitting back. However, these assumptions have not been tested experimentally.

Neural Bases of Costly Punishment

Neuroscientific studies provide further insights into the roots of pro-social punishment. However, no neural studies on antisocial punishment are available yet. These studies provide evidence that the decision to punish is influenced by a complex interplay of cognitive and emotional processes involving various parts of the brain.

Emotional factors: Among possible emotional factors, satisfaction has by far received the most attentions (although other emotions have been discussed, see e.g. Strobel et al., [2011] on disgust). de Quervain et al. (2004) found that (effective, as opposed to symbolic) punishment in a trust game activated the punisher's dorsal striatum, a region of the brain that has been implicated in the integration of behavioural and reward information into a goal-directed mechanism, as well as the processing of rewards gained from making a decision. Subjects who experienced higher activation in the caudate nucleus (part of the dorsal striatum) when punishing defectors were also willing to expend a greater share of their endowment on punishment when it was costly.

de Quervain et al.'s (2004) study provides evidence that people derive satisfaction from punishment behaviour that violates the fairness norm. The researchers tested two hypotheses about the underlying mechanism: a) stronger punishment causes stronger activation of the caudate nucleus, or b) subjects expecting higher satisfaction from punishing a norm violation could be willing to incur higher costs for punishing. They found support for the latter hypothesis, i.e. the dorsal striatum reflects expected satisfaction from punishment. People who expect to be satisfied by punishing a defector

will go to greater lengths to do so.

The emotional processes involved in making decisions about pro-social punishment are largely similar independent of personal involvement and the effectiveness of punishment. However, Strobel et al. (2011) found that the nucleus accumbens was more involved in the first-person treatment (as compared to the third-party treatment), whereas the nucleus caudatus was more involved when punishment had a strong effect (as compared to a weak effect). This, however, is only a quantitative effect, i.e. a common motivational mechanism might underlie the punishment of defectors independent of personal involvement and the effectiveness of punishment.

In Strobel et al.'s (2011) study, the insula was activated more strongly when subjects decided to punish defectors. This area of the brain has been assumed to reflect representations of emotional states and norm violations. This is further evidence of the role of emotions, and in particular negative ones, as the authors assume, in making punishment decisions.

Cognitive factors: Neurological studies also point towards an involvement of cognitive processes in pro-social punishment, at least when punishment is costly. de Quervain et al. (2004) found that under costly punishment conditions, the ventromedial prefrontal and the medial orbitofrontal cortex were activated more strongly (as compared to free punishment). The ventromedial prefrontal cortex has been implicated in the integration of multiple cognitive processes in the pursuit of behavioural goals, and the medial orbitofrontal cortex is associated with difficult choices that require the coding of reward value.

Strobel et al. (2011) were unable to replicate de Quervain et al.'s (2004) findings with regard to the orbitofrontal cortex, but found increased activity dorsolateral prefrontal cortex as well as the anterior cingulate cortex. They point out the role of the dorsolateral prefrontal cortex in the implementation of cognitive control and the temporal organisation of goal-directed action. The anterior cingulate cortex has been linked to emotional and cognitive conflict, where it is likely involved in monitoring the need to exert control to promote behavioural adjustments. These activations of the prefrontal, orbitofrontal, and anterior cingulate cortex point towards a difficult cognitive task posed by the complex decision problem that is costly punishment.

Genetic Bases of Costly Punishment

Few studies have examined genetic factors that might be responsible for within- and across-population variation in punishment behaviour. However, Strobel et al. (2011) investigated a genetic variation of dopamine function. This involves a polymorphism of the gene encoding the dopamine-degrading enzyme catechol-O-methyltransferase (COMT). They found that a variation of COMT linked to higher synaptic dopamine availability was associated with higher punishment-related activation of the nucleus accumbens. As the authors argue, this might bias the integration of input signals from dorsolateral prefrontal cortex, anterior cingulate cortex, and insula due to a higher reward anticipation associated with the decision to punish norm-violating behaviour.

A Model of Punishment Decision-Making

Emotional, cognitive and motivational processes impact decision-making about pro-social punishment. These factors interact in a complex way. Based on their findings, Strobel et al. (2011) have outlined a model for these interactions, which, although untested with regard to causalities, provides an intriguing structure:

"[I]nsular representations of negative emotional states due to norm violations would provide a bias signal, which interferes with signals of immediate individual financial reward if no punishment is exerted. The resulting behavioural conflict, monitored and signalled by the ACC, would result in DLPFC-mediated implementation of cognitive control, which would impact on striatal integration of the input signals in favour of the decision to punish, given that more future reward would be anticipated following such behaviour due to learned contingencies between norm-conform or norm-enforcing behaviour and social reward."

This model integrates cognitive, emotional, and motivational processes implicated in decision-making about pro-social punishment. With regard to the evidence presented in this paper, however, further integration of environmental factors and an enhancement to include antisocial punishment would be necessary to make this a comprehensive

model.

Conclusion

Mainstream economic theory, and indeed much of the social sciences, build upon expected utility theory, which holds that people make decisions based on rational self-interest. Evidence from economic games, however, show that these assumptions are frequently violated. In particular, humans across different societies and in a variety of settings have shown willingness to expend money on pro-social punishment, even though this violates rational self-interest.

Behavioural economists have argued that pro-social punishment is a phenomenon of other-regarding (as opposed to self-interested) preferences. Fehr, Fischbacher, and Kosfeld (2005) argue that pro-social punishment is, indeed, a rational path of action for individuals with social preferences. They interpret activation in the striatum, a reward centre of the brain, associated with pro-social punishment (as well as mutual cooperation) as "non-pecuniary utility", i.e. set it equal to monetary utility derived from an interaction. This interpretation implies that economic models would need to be enhanced to include non-pecuniary utility derived from the mode of interaction, rather than the exchange itself, in individuals' total utility; however, the assumption of rationality would be held up.

The interpretation put forward by Fehr, Fischbacher, and Kosfeld (2005) is consistent with the theory of strong reciprocity put forward by Gintis (2000). This theory holds that humans have are "predisposed to cooperate with others and punish non-cooperators, even when this behaviour cannot be justified in terms of extended kinship or reciprocal altruism" (ibid.). The evolution of strong reciprocity has been cast in terms of gene-culture co-evolution, i.e. a form of group selection (e.g. Nowak & Sigmund, 2005).

Hagen and Hammerstein (2006) have criticized the cognitive model underlying behavioural economists' interpretations of economic experiments. They argue that strong reciprocity theory mistakes the brain for a "utility-maximizing machine". Instead, seeing the brain as modular, not one, but a great variety of mechanisms could be at work in making decisions in economic games. "[P]layers might be using a variety of frames to

guide their choices. Sometimes the frame might be a pan-human frame [...] evolved by natural selection. Other times it might be a social or economic institution that was acquired culturally or by individual learning," Hagen and Hammerstein posit. This interpretation would accommodate the currently puzzling within- and across-population variability in the extent of pro-social punishment.

The decision to incur costs in order to pro-socially punish defectors is a complex cognitive task. It is guided in part by emotions, which might be phenomena of social preferences. However, cultural and environmental factors also impact this decision-making process. So far, research into pro-social punishment has been concentrated with behavioural economists, as well as, to some extent, theoretical biologists and economic anthropologists. There is, however, so far little involvement from cognitive psychologists. Building comprehensive models that include the various processes involved will thus be one of the main tasks for the future.

References

- de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*, 1254-1258.
- Fehr, E., Fischbacher, U., & Kosfeld, M. (2005). Neuroeconomic Foundations of Trust and Social Preferences: Initial Evidence. *The American Economic Review*, *95*(2), 346-351.
- Fehr, E. & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*, 980–994.
- Fehr, E. & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137-140.
- Fehr, E. & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics*, *114*(3), 817-868.
- Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society: Biological Science*, *365*, 2651-2661.
- Gächter, S., Renner, E., & Sefton, M. (2008). The Long-Run Benefits of Punishment. *Science*, *322*, 1510.
- Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical*

- Biology*, 206, 169-179.
- Goldstein, E. B. (2011). *Cognitive Psychology*. Florence, KY: Wadsworth Cengage Learning.
- Hagen, E. & Hammerstein, P. (2006). Game theory and human evolution: a critique of some recent interpretations of experimental games. *Theoretical Population Biology*, 69(3), 339-348.
- Henrich et al. (2006). Costly Punishment Across Human Societies. *Science*, 312, 1767-1770.
- Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial Punishment Across Societies. *Science*, 319, 1362-1367.
- Nowak, M. A. & Sigmund K. (2005). Evolution of indirect reciprocity. *Nature*, 437, 1291-1298.
- Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants With and Without a Sword: Self-Governance is Possible. *American Political Science Review*, 86(2), 404-417.
- Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *NeuroImage*, 54(1), 671-680.