2-2012

The Restorative Logic of Punishment: Another Argument in Favor of Weak Selection

Nicolas Baumard
University of Pennsylvania

Recommended Citation

Baumard, N. (2012). The Restorative Logic of Punishment: Another Argument in Favor of Weak Selection. *Behavioral and Brain Sciences, 35* (1), 17-18. http://dx.doi.org/10.1017/S0140525X11001166

The Restorative Logic of Punishment: Another Argument in Favor of Weak Selection

Abstract

Strong reciprocity theorists claim that punishment has evolved to promote the good of the group and to deter cheating. By contrast, weak reciprocity suggests that punishment aims to restore justice (i.e., reciprocity) between the criminal and his victim. Experimental evidences as well as field observations suggest that humans punish criminals to restore fairness rather than to support group cooperation

Disciplines

Criminology and Criminal Justice | Psychology | Social Psychology and Interaction

are predicted to be adaptive *on average* outside the laboratory; for example, if being nice invites reciprocation. People bring these psychological mechanisms with them into the laboratory, where the behaviour produced may or may not still be adaptive on average (Barclay, 2011; West et al. 2011). "Maladaptive" behaviour can persist despite repeated anonymous encounters, as long as the same proximate psychological mechanisms are repeatedly triggered (e.g., anger, desire for fairness, empathy). However, this would say little about the ultimate function that those mechanisms serve outside the laboratory. Too much ink has been spilled by researchers who do not realize that their colleagues are simply addressing a different level of analysis.

On a completely different note, Guala makes a useful distinction between wide and narrow readings of the experimental evidence, and what each reading implies. Wide interpretations can clearly be taken too far: If punishment (or any other phenomenon) supports cooperation in the lab, it does not necessarily mean that this is what supports it outside the lab. However, I would caution against hasty abandonment of such wide interpretations. Sometimes laboratory experiments use controlled conditions to test whether a proposed mechanism *could* support punishment. At other times, such experiments test the validity of theories of human behaviour (Mook 1983): If a predicted phenomenon cannot be found in the lab under ideal controlled conditions, then we must either reject or revise any theory that relies on that phenomenon (see, e.g., the lack of punishment towards non-punishers in Kiyonari & Barclay 2008). If successful, do these findings need confirmatory non-laboratory observations with real-world phenomena? Absolutely. Convergent evidence is crucial in all scientific enterprises, and the laboratory and the field have their own respective strengths and weaknesses. As such, we should all strongly support the call for collaborations across disciplines and between the lab and the field. Guala's target article has clearly shown that the punishment literature needs more of this, and for that it should be commended.

ACKNOWLEDGMENTS
I thank Daniel Krupp for discussions and the Social Sciences and Humanities Research Council of Canada (SSHRC) for support.

The restorative logic of punishment: Another argument in favor of weak selection

doi:10.1017/S0140525X11001166

Nicolas Baumard
Philosophy, Politics, and Economics Program, University of Pennsylvania, Philadelphia, PA 19104.
nbaumard@gmail.com
https://sites.google.com/site/nicolasbaumard/Home

Abstract: Strong reciprocity theorists claim that punishment has evolved to promote the good of the group and to deter cheating. By contrast, weak reciprocity suggests that punishment aims to restore justice (i.e., reciprocity) between the criminal and his victim. Experimental evidences as well as field observations suggest that humans punish criminals to restore fairness rather than to support group cooperation.

As Guala rightly notes, there is very little evidence that punishment plays a role in the stabilization of cooperation in small-scale societies. On the other hand, as he also notes, it is difficult to totally rule out the strong view of punishment as it is complicated to precisely assess the costs of punishment in the field (Are there really no costs in punishing others? Aren't there many hidden benefits for the individual who punish? etc.). There is, however, another way to disentangle the two views of punishment, namely, the forms that punishments take. Indeed, the two theories – the weak and the strong – make different predictions regarding the logic of punishment.

Group selection theory holds that punishment aims to promote the good of the group by sustaining cooperation and preventing cheating (Boyd et al. 2003; Fehr & Gächter 2002; Henrich & Boyd 2001). This implies that punishment should be calibrated to deter crimes and render them non-advantageous. Here, group selection parallels the utilitarian doctrine of punishment, which contends that punishment should be used to deter crimes and maximize the good of society (Polinsky & Shavell 2000; Posner 1983). The utilitarian theory of punishment holds, for instance, that the detection rate of a given crime and the publicity associated with a given conviction are relevant factors in assigning punishments. If a crime is difficult to detect, the punishment for that crime ought to be made more severe in order to counterbalance the temptation created by the low risk of getting caught. Likewise, if a conviction is likely to get a lot of publicity, a law enforcement system interested in deterrence should take advantage of this circumstance by "making an example" of the convict with a particularly severe punishment, thus getting a maximum of deterrence for its punishment.

By contrast, individual selection predicts a "restorative" or "retributive" logic for punishment (Baumard 2011). Restorative logic holds that punishment aims to restore justice between the criminal and the victim – either by harming the criminal or by compensating the victim. In intuitive terms, people are punished because they "deserve" to be punished, and not because punishing them would be useful for the society at large.

This restorative logic is a direct consequence of the way cooperation has evolved among humans (Baumard 2010a; Trivers 1971). Indeed, human beings belong to a highly cooperative species and get most of their resources from collective actions, solidarity, exchanges, and so forth. (Gurven 2004; Hill & Kaplan 1999). In the ancestral environment, individuals were in competition to be recruited for the most fruitful ventures, and it was vital to share the benefits of cooperation in a mutually advantageous manner. If individuals took a bigger share of the benefits, their partners would leave them for more interesting partners. If they took a smaller share, they would be exploited by their partners who would receive more than what they had contributed to produce. This competition to attract cooperative partners is thus likely to have led to selection for a "sense of fairness," a cognitive device that motivates individuals to share the costs and benefits of social interaction in an impartial way (André & Baumard 2011). If cooperation is based on fairness, then crimes create an unfair relationship between the criminal and her victim, and people have the intuition that the criminal ought to compensate the victim or to be punished in order to restore justice.

It is worth mentioning that this theory does not mean that punishment should be absent in human societies. As Guala notes, modern societies have found many institutional ways to reduce the costs of punishments. Although these institutions are absent in smaller societies, justice can still be restored by individuals seeking to retaliate. Retaliation is indeed advantageous from an individual perspective and can indeed be found in many nonhuman species (Clutton-Brock & Parker 1995). As Evans-Pritchard noted, in societies where there is no penal system, "self-help, with some backing of public opinion, is the main sanction" (Evans-Pritchard 1940/1969, p. 169).

In this kind of situations, selfish and moral motives converge: The victim (or his allies) attacks the criminal to signal his strength and gains a reputation as someone who cannot be attacked without risk; and by doing so, he also punishes the wrongdoer by allowing justice to be done. In line with this idea, people in small-scale societies distinguish between legitimate (and proportionate) retaliation and illegitimate (and disproportionate) retaliation (von Fürer-Hameindorf 1967; Miller 1990). Retaliation is thus clearly limited by moral concerns: within the group, it has to be proportionate to the prejudice. As the *Lex Talionis* says, "an eye for an eye, a tooth for a tooth," but no more.

Individual selection thus clearly predicts some kind of punishment, and, more importantly, it predicts that punishments should aim toward a specific goal (restoring fairness) that differs from the utilitarian goal predicted by group selection (preventing wrongdoing). Experimental studies, relying on a variety of methodologies, suggest that punishments fit individual selection more than group selection. Indeed, when people punish harmdoers, they generally respond to factors relevant to a retributive theory of punishment (magnitude of harm, moral intentions) and ignore factors relevant to the group selection theory (likelihood of detection, publicity, likelihood of repeat offending) (Baron et al. 1993; Baron & Ritov 2008; Carlsmith et al. 2002; Darley et al. 2000; Glaeser & Sacerdote 2000; Sunstein et al. 2000).

In line with these results, field observations have extensively demonstrated that, in keeping with the prediction, the level of compensation in stateless societies is directly proportional to the prejudice inflicted to the victim: For example, the wrongdoer owes more to the victim if the wrongdoer has killed a family member or eloped with a wife than if he has stolen animals or destroyed crops (Hoebel 1954; Howell 1954; Malinowski 1926). To conclude, punishment does not seem to be a group adaptation. It follows the logic of fairness rather than the interests of the group.

Reciprocity and uncertainty

doi:10.1017/S0140525X11001178

Yoella Bereby-Meyer

Department of Psychology, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel.

Yoella@bgu.ac.il   www.YoellaBerebyMeyer.com

Abstract: Guala points to a discrepancy between strong negative reciprocity observed in the lab and the way cooperation is sustained "in the wild." This commentary suggests that in lab experiments, strong negative reciprocity is limited when uncertainty exists regarding the players' actions and the intentions. Thus, costly punishment is indeed a limited mechanism for sustaining cooperation in an uncertain environment.

Strong reciprocity is the behavioral predisposition to cooperate conditionally on others' cooperation and to punish violations of cooperative norms even at a net cost to the punisher (Fehr & Gintis 2007). The phenomenon has been the subject of considerable research in the last few decades (e.g., Camerer 2003; Fehr & Gächter 2000b; Rabin 1993), and its existence is well established.

In the target article, Guala points to a discrepancy between the strong negative reciprocity that is observed in the lab and the way cooperation is sustained "in the wild." Specifically, he suggests that there is no indication for costly punishment in the wild. This claim gives rise to the question as to what extent one can predict actual behavior in real-life situations from behavior in the very artificial and contrived laboratory setting. The author suggests that behavior in the laboratory with respect to strong negative reciprocity does not really reflect behavior in real life. However, the matter may actually be somewhat more complex than it seems. Even if there is no strong negative reciprocity in the real world, this may still be in line with the results from laboratory studies. One simply has to make sure that the laboratory studies capture crucial characteristics of the real world.

One of the main properties of real-world situations is some degree of uncertainty (which does not usually exist in laboratory studies and particularly in those the author referred to). In many real-life social dilemmas people face uncertainty of two types: (1) environmental uncertainty, which is uncertainty regarding aspects of the dilemma (e.g., the size of the common resource); and (2) social uncertainty, which is uncertainty regarding the other group members' choices (Messick et al. 1998). Moreover, outcomes may be determined probabilistically.

For negative reciprocity to occur, accurate knowledge regarding the actions and the intentions of the players is important. If uncertainty exists, it will be difficult to determine whether the action and the outcome were the result of violations of cooperative norms. While most laboratory experiments have dealt with situations that are certain, a number of studies have introduced some degree of uncertainty into situations in which negative reciprocity is possible. These studies consistently show that uncertainty lowers the tendency towards negative reciprocity.

Most of the evidence for strong negative reciprocity was observed in the Ultimatum Game (UG). In research on the UG, responders are very likely to reject offers that are less than 30% of the cake (e.g., Güth et al. 1982). By rejecting the offer, the responder gives up a possible gain; and thus the finding is interpreted as evidence for responders' willingness to pay a cost to punish the proposers they perceive as acting unfairly, even if they will never meet the proposers again. In the classic experiment and in most following ones, the size of the pie to be divided is common knowledge. As was noted by Croson (1996), this assumption is unrealistic.

Several experiments investigated UGs with one-sided uncertainty on the part of the responder. Typically, in these experiments proposers know the exact amount of money to be divided, and responders either know nothing at all or know the probability distribution of possible amounts.

In most studies responders accepted lower offers when they did not know the size of the pie and when the lack of information was common knowledge. Proposers, in turn, did not hesitate to exploit this behavior and offered little when the amount to divide was large (e.g., Croson 1996; Mitzkewitz & Nagel 1993; Rapoport et al. 1996). Recently Gehrig et al. (2007) studied a UG with a different source of uncertainty. In their game the responder knows the pie size but not the offer when deciding whether to accept or reject (i.e., has imperfect information). Responders never reject in this game, even when they anticipate low offers.

Under both types of uncertainty responders seem to give proposers the benefit of the doubt: Because a low offer could be fair if the pie is small or the yet-unknown offer could eventually be fair, rejecting the offer would mean punishing the proposer unfairly. Consequently, with uncertainty lower offers are more likely to be accepted. This behavior is strong evidence that rejections in the UG are an expression of preference when responders do know the proposer's payoff (Camerer 2003); and therefore the ability to generalize these preferences to situations with uncertainty is limited.

Uncertainty also affects reciprocity in repeated interactions. The ability of reciprocity to sustain cooperation in the long run, and specifically in the iterated Prisoner's Dilemma, was demonstrated by Axelrod's (1984) well-known computer tournaments. Later it has been shown that cooperation is much more difficult to maintain if there is uncertainty regarding players' actions, that is, if there is random error either in choosing actions or in monitoring others' actions (see, e.g., Axelrod & Dion 1988; Bendor 1993; Green & Porter 1984; Sainty 1999). That is, if actions are noisy, a player does not know whether another player's defection was an error or an intended choice, and strategies involving reciprocation (e.g., tit for tat) can break down. But even if players can monitor others' past actions perfectly in a repeated Prisoner's Dilemma game, if payoffs are noisy, players learn to cooperate much less (e.g., Bereby-Meyer & Roth 2006; Kunreuther et al. 2009).

Hence, the fact that Guala in his analysis of real-world situations did not find evidence for strong negative reciprocity does not necessarily imply that results from laboratory studies cannot predict reciprocity behavior in the real world. Instead, one can perhaps conclude that in situations with uncertainty,