worse on subsequent tests of their physical and mental capacities (e.g., they solved fewer anagrams) then participants who did not ostracize a confederate (Ciarocco et al. 2001). Even a seemingly innocuous form of punishment, gossip, is not without social costs: People who gossip negatively about others are less trusted and are more prone to negative reputations than those who do not gossip, even when controlling for the frequency of gossip (Turner et al. 2003). Finally, verbal reproach is also socially costly. Whistle-blowers who speak up against illegal behaviors perpetrated by employees of their organizations are susceptible to retaliation (e.g., negative performance evaluations, ostracism, dismissal) from members of the organization (Miceli et al. 2008; Near & Miceli 1995; Rothschild & Miethe 1999). Indeed, the prevalence of retaliation against whistle-blowers has led to the passage of legislation in the United States and other countries to attempt to protect whistle-blowers (see Miceli et al. [2008] for a review). Other examples of the costs of verbal reproach abound: Whites such as Viola Liuzzo who protested racial discrimination and segregation during the Civil Rights Movement in the United States suffered physical harm, reputational and material costs, and even death (Stanton 2000). In sum, ostracism, gossip, and verbal reproach can all be psychologically or socially costly forms of punishment. Although many factors likely influence whether these costs are experienced in any given situation, we simply highlight that punishers sometimes incur such costs.

Beyond these psychological and social costs, there is also anecdotal evidence of material costs associated with punishing "in the wild," such as when individuals or groups choose to boycott an organization. For example, the Dean and faculty at Vermont Law School denied military recruiters access to their campus facilities for many years because they opposed the military's "Don't Ask, Don't Tell" policy that prevents those who are openly gay, lesbian, or bisexual from serving in the military. As a result, the military had a difficult (but not impossible) time recruiting Vermont Law students, and the school forfeited approximately $500,000 in federal funding annually (Sanchez 2005).

Given the various types of costs we have reviewed, it is worth noting that empirical evidence supports Guala's speculation that people's emotions or motivations might lead them to punish even when it is against their immediate self-interest. Psychological research demonstrates that people's desires to punish are driven primarily by retribution, such that people punish to see the offenders suffer in a manner proportionate to their wrongdoing, even if the punishment will not effectively deter future transgressions (see Carlsmith & Darley [2008] for a review). In other words, people may punish to satisfy their retributive desires, even when it is costly to do so.

In conclusion, Guala dismisses non-material costs by claiming that they are not very costly or that they are not relevant to arguments of group fitness. In contrast, we argue that broadening the definition of costs to include social and psychological costs can help to inform the debate about whether there is evidence of costly punishment "in the wild."

Proximate and ultimate causes of punishment and strong reciprocity

doi:10.1017/S0140525X11001154

Pat Barclay

Department of Psychology, University of Guelph, East Guelph, Ontario, N1G 2W1, Canada.

barclayp@uoguelph.ca

http://www.uoguelph.ca/nacs/page.cfm?id = 229

Abstract: While admirable, Guala's discussion of reciprocity suffers from a confusion between proximate causes (psychological mechanisms triggering behaviour) and ultimate causes (evolved function of those psychological mechanisms). Because much work on "strong reciprocity" commits this error, I clarify the difference between proximate and ultimate causes of cooperation and punishment. I also caution against hasty rejections of "wide readings" of experimental evidence.

Guala reviews a number of interesting field studies that speak against the importance of punishment in maintaining cooperation. This is important because there is an abundance of laboratory research on punishment and cooperation which has outstripped the research in real-world settings. Underlying much of Guala's discussion of reciprocity and punishment, however, there lies confusion over proximate causation and ultimate causation. Confusion over these levels of analysis is not only present in Guala's target article, but is endemic to the entire field of cooperation and is particularly pronounced in the discussion of "strong reciprocity." This weakens Guala's arguments. In particular, it results in unwarranted statements against so-called weak reciprocity. As such, this topic requires clarification.

Any behaviour, including cooperation and punishment, can be explained at four different levels of analysis (Tinbergen 1968). Proximate causes include: (1) the psychological mechanisms that trigger behaviour (e.g., emotions, cognitions); and (2) the developmental processes that cause those psychological mechanisms to arise within an individual's lifetime (e.g., "innate" behaviour, learning, internalization of cultural norms). Ultimate causes include: (3) the evolutionary forces (e.g., reciprocity, mutualism, costly signalling) that result in those psychological mechanisms existing instead of other possible psychologies; and (4) the evolutionary history of those mechanisms and when they arose in our lineage (e.g., unique to humans, shared with other primates). These four levels of analysis – mechanism, development, function, and phylogeny – are complementary, not mutually exclusive. A complete explanation of any phenomenon requires an answer at each level.

An example can help clarify the proximate and ultimate causes of cooperative behaviour. Suppose that I genuinely value your welfare and I help you without any ulterior motives. If my action causes you to genuinely care about me, you will be more likely to help me when I need it, even when you anticipate no benefits for doing so. If I happen to find out, then your actions will cause me to value your welfare more and help you more often, and so on. The reciprocity in this example is not "weak": both of us unselfishly reciprocate "altruistic" acts. Both of us do benefit from helping each other, but neither one intended to benefit, and neither of us requires any foresight of the consequences. Helping can be altruistic from a proximate psychological perspective, but from an ultimate (functional) perspective it is advantageous to possess such a psychology. Thus, contrary to Guala's assumption, biologists do not assume *psychological* self-interest. To paraphrase Dawkins (1976/2006): The genes are selfish, but this doesn't mean the person is. One can make a similar argument with punishment: I may punish you because I am angry (proximate cause), and this may result in me receiving more future cooperation from you (potential ultimate cause of punitive sentiment), but this does not mean that my punishment was motivated by a desire for your cooperation.

Guala uses terms like "strong" and "weak" reciprocity, which are often misleading because they often conflate the proximate psychological mechanisms with the ultimate functional reasons for why those psychological mechanisms exist (Barclay 2010; West et al. 2007b). By itself, "strong reciprocity" is merely a description of behaviour, that is, the supposed tendency of people to cooperate, reward cooperators, and punish cooperators, even when there are no apparent benefits for doing so. The goal is to discover – at *all* levels among levels of analysis – why this tendency exists (if indeed it does). So-called theories of "weak reciprocity" are often theories about the ultimate function of cooperative and punitive sentiment, not theories about what specifically that sentiment is or how it develops. People possess certain emotions and psychological mechanisms which

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.