Title: Examining punishment at different explanatory levels.
Authors: dos Santos M., Wedekind C.

Journal: Behavioral and Brain Sciences,
Year: 2012

Volume: 35(1)

Pages: 23-24

DOI: 10.1017/S0140525X1100121X

Examining punishment at different explanatory levels

Miguel dos Santos & Claus Wedekind

Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland.

ABSTRACT

Experimental studies on punishment have sometimes been overinterpreted not only for the reasons Guala lists but also because of a frequent conflation of proximate and ultimate explanatory levels that Guala's review perpetuates. Moreover, for future analyses we may need a clearer classification of different kinds of punishment.

When explaining behavioral decisions, it is important to distinguish between different explanatory levels, especially between proximate (mechanistic) and ultimate (evolutionary) explanations (Tinbergen, 1963). Proximate explanations of a given behavior deal with questions about its ontogeny (e.g. how does the behavior change with age and experience) or about its causation, i.e. the physiological, molecular, and cognitive mechanisms underlying the behavior and the stimuli that elicit it. Ultimate explanations either deal with questions about the phylogeny of the behavior (e.g. how does it compare with similar behaviors in related species) or its adaptive value (e.g. what is its impact on the individual's survival and life-time reproductive success).

The concept of weak reciprocity, as defined in Guala (2011), is an attempt to explain the adaptive value of cooperation and punishment because it concentrates on the fitness benefits one could get from cooperating, defecting, or punishing (Trivers, 1971; Alexander, 1974). This concept is restricted to one explanatory level only. In contrast, strong reciprocity mixes different explanatory levels: it uses proximate arguments to explain ultimate problems (Fehr & Gächter, 2002; Fehr & Fischbacher, 2003, 2004; Fehr & Rockenbach, 2003; Gintis et al., 2003; Bowles & Gintis, 2004). Strong reciprocity is, for example, called a "… predisposition to reward others for cooperative, norm-abiding behaviours" and "… a propensity to impose sanctions on others for norm violations" (Fehr & Fischbacher, 2003, p. 785). Such a definition clearly relates to the causal mechanisms of cooperation and punishment. But the concept is then frequently used as to answer ultimate (evolutionary) questions, for example in Bowles & Gintis (2004, p.17):"… cooperation is maintained because many humans have a predisposition to punish those who violate group-beneficial norms". Such a mixing up of different explanatory levels can, from an evolutionary point of view, easily lead to overinterpretations of proximate patterns (Hagen & Hammerstein, 2006; Sigmund, 2007; West et al., 2007, in press; Rankin et al., 2009). For example, punishment that can be observed in anonymous one-shot interactions seems truly altruistic and was interpreted as such in Fehr & Gächter (2002). However, until very recently, humans lived in groups where anonymous one-shot interactions were probably very rare, i.e. such interactions are most probably not the context in which human punishment has

evolved. If studied within a more natural social context, human punishment may ultimately be self-interested.

As discussed in Guala (2011), explaining punishment from an evolutionary point of view requires determining the costs and benefits of punishment. In line with weak reciprocity models, recent studies have shown that punishment can lead to long-term net benefits and hence be evolutionarily stable when punitive actions contribute to a punishment reputation (Hilbe & Sigmund, 2010; dos Santos et al., 2011). Under such conditions, the immediate costs of punishment can be outweighed by the benefits a punisher receives later because of his/her punishment reputation. Experimental studies that ignore the possible effects of a punishment reputation can therefore easily produce artifacts (Hagen & Hammerstein, 2006).

We also believe that the term "punishment" is currently used too broadly in the literature on cooperation. If "punishment" is the subtraction of resources from free-riders in order to reduce the frequency of further free-riding, there are at least three different kinds of punishment that may need to be distinguished both for ultimate and proximate analyses. Many of these analyses deal with what could be called "simple costly punishment", i.e. punishers pay a cost to induce a cost on the punished (Fehr & Gächter, 2000; Rockenbach & Milinski, 2006; Dreber et al., 2008; Rand et al., 2009; Wu et al., 2009). Another form of punishment could be called "punishment by taking something away" (e.g. Cephu's example in Guala, 2011). Here, the punisher takes something from the punished in order to induce a cost to the punished. Regardless of whether the punisher thereby experiences an immediate reduction of the own welfare or not, "punishment by taking something away" and the upper "simple costly punishment" are likely to differ in their cost-benefit ratios (relevant for ultimate analyses) and may involve, for example, different kinds of emotions (relevant for proximate analyses). A third category could be called "punishment by refusal". The punisher then punishes by refusing to cooperate with the punished in a repeated game like, for example, an iterated Prisoner's Dilemma (Fudenberg et al., 1994). The examples of ostracism discussed in Guala relate to this kind of punishment. Such defection may typically be a reaction to non-provoked defection and could be called "punishment" if it reduces the income of the punished (i.e. his/her benefits from what would otherwise be cooperative interactions) in order to possibly improve the punisher's long-term benefits from future cooperative interactions with a refined punished or with others. This third kind of punishment could be immediately costly for the punisher, for example, if it delays the resumption of beneficial mutual cooperation. Such immediate costs would have to be compensated on the long run in order to maintain "punishment by refusal" as an evolutionary successful behavioral strategy. However, a possible alternative function of defection in response to defection may be to simply avoid the losses of anticipated further defection (e.g. avoiding the sucker's payoff in the Prisoner's Dilemma). It is probably not useful to call this later form of defection "punishment" if it usually does not ultimately increase the level of cooperation within a group or directly with the defector (from an ultimate point of view), or if it is just a precautionary measure to avoid further losses (from a proximate point of view). Therefore, purely punitive actions may not always be easy to identify. Multidisciplinary approaches that carefully exploit the specific advantages of proximate and ultimate analyses are therefore often necessary to better understand human behavior.

REFERENCES

Alexander, R. D. (1974). The evolution of social behaviour. *Annual Review of Ecological Systematics*, 5, 325-383.

Bowles, S. & Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, 65, 17-28.

dos Santos, M., Rankin, D. J. & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B-Biological Sciences*, 278, 371-377.

Dreber, A., Rand, D. G., Fudenberg, D. & Nowak, M. A. (2008). Winners don't punish. *Nature*, 452, 348-351.

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

Fehr, E. & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25, 63-87.

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

Fehr, E. & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415, 137-140.

Fehr, E. & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422, 137-140.

Fudenberg, D., Levine, D. & Maskin, E. (1994). The Folk Theorem with Imperfect Public Information. *Econometrica*, 62, 997-1039.

Gintis, H., Bowles, S., Boyd, R. & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, 24, 153-172.

Hagen, E. H. & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology*, 69, 339-348.

Hilbe, C. & Sigmund, K. (2010). Incentives and opportunism: from the carrot to the stick. *Proceedings of the Royal Society B-Biological Sciences*, 277, 2427-2433.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D. & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science*, 325, 1272-1275.

Rankin, D. J., dos Santos, M. & Wedekind, C. (2009). The evolutionary significance of costly punishment is still to be demonstrated. *Proceedings of the National Academy of Sciences USA*, 106, E135-E135.

Rockenbach, B. & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature*, 444, 718-723.

Sigmund, K. (2007). Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology & Evolution*, 22, 593-600.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift fur Tierpsychologie*, 20, 410-433.

Trivers, R. L. (1971). Evolution of reciprocal altruism. *Quarterly Review of Biology*, 46, 35-57.

West, S. A., El Mouden, C. & Gardner, A. (In press). 16 common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*,

West, S. A., Griffin, A. S. & Gardner, A. (2007). Evolutionary explanations for cooperation. *Current Biology*, 17, R661-R672.

Wu, J. J., Zhang, B. Y., Zhou, Z. X., He, Q. Q., Zheng, X. D., Cressman, R. & Tao, Y. (2009). Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences USA*, 106, 17448-17451.