Title: Human cooperation based on punishment reputation.
Authors: dos Santos M., Rankin D.J., Wedekind C.

Journal: Evolution
Year: 2013

Volume: 67(8)

Pages: 2446-2450

DOI: 10.1111/evo.12108

Human cooperation based on punishment reputation

Miguel dos Santos[1], Daniel J. Rankin[2,3] & Claus Wedekind[1*]

[1] Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland.
[2] Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.
[3] Swiss Institute of Bioinformatics, Quartier Sorge Bâtiment Génopode, 1015 Lausanne, Switzerland.

* Tel: +41 21 6924250; Fax: +41 21 6924265; E-mail: claus.wedekind@unil.ch

Running title: Punishment and reputation

ABSTRACT

The threat of punishment usually promotes cooperation. However, punishing itself is costly, rare in non-human animals, and humans who punish often finish with low payoffs in economic experiments. The evolution of punishment has therefore been unclear. Recent theoretical developments suggest that punishment has evolved in the context of reputation games. We tested this idea in a simple helping game with observers and with punishment and punishment reputation (experimentally controlling for other possible reputational effects). We show that punishers fully compensate their costs as they receive help more often. The more likely defection is punished within a group, the higher the level of within-group cooperation. These beneficial effects perish if the punishment reputation is removed. We conclude that reputation is key to the evolution of punishment.

Keywords: experimental game theory, punishment, indirect reciprocity
Data to be archived in Dryad

1. INTRODUCTION

Punishment of non-cooperators often promotes cooperation in humans (Yamagishi 1986; Fehr and Gächter 2000, 2002; Rockenbach and Milinski 2006; Sigmund 2007) and other animals (Bshary and Grutter 2005; Raihani et al. 2012). When punished, non-cooperators are more likely to cooperate not only with the punisher but also with other social partners. Although usually conferring benefits on the punisher's social group on the long run (Yamagishi 1986; Fehr and Gächter 2000; Gächter et al. 2008; Wu et al. 2009), punishing defectors is costly to the punisher. Thus, punishment has often been perceived as truly altruistic (Fehr and Gächter 2002; Barclay 2006) and its evolutionary significance as puzzling (Dreber et al. 2008; Rankin et al. 2009; Tao et al. 2009; Wu et al. 2009; dos Santos and Wedekind 2012).

Reputation within a social group, for example by using an individual's "image scoring" or "standing", has been proposed as a potentially important mechanism to explain cooperation in humans (Nowak and Sigmund 1998; Wedekind and Milinski 2000; Milinski et

al. 2002; Wedekind and Braithwaite 2002; Nowak and Sigmund 2005). Several findings suggest that reputation could also be important for the evolution of punishment. For example, humans playing an ultimatum game reject lower offers when they know that others will learn about their acceptance threshold (Fehr and Fischbacher 2003). They also seem to be more likely to punish as the number of observers increases (Kurzban et al. 2007), and punishers in public goods games are perceived as more "group focused" and receive more monetary benefits in a consecutive trust game (Barclay 2006). However, previous experimental studies on human punishment have either not allowed for reputational effects, or the kind of reputation that could built up was not clearly defined because reputational effects of punitive actions were potentially confounded with reputational effects of cooperative actions (Fehr and Gächter 2000; Rockenbach and Milinski 2006; Dreber et al. 2008; Gächter et al. 2008; Rand et al. 2009; Ule et al. 2009; Wu et al. 2009).

Theory predicts that individuals benefit from taking the likelihood of being punished into account (Brandt et al. 2003; Gardner and West 2004; Hilbe and Sigmund 2010). Punishing defection can then, on the long run, be advantageous to the punisher and hence evolve if it builds up a punishment reputation (Hilbe and Sigmund 2010; dos Santos et al. 2011; Hilbe and Traulsen 2012). This leads to two key predictions: (i) humans use reputation to discriminate between non-punishers and punishers and are more cooperative to the latter, and (ii) the immediate costs of punishment are compensated over time by the additional cooperation the punisher receives from punished and observers. We tested these predictions in a helping game with observers and with the option of punishment. Experimentally controlling for potential reputational effects of cooperation and defection allowed us to specifically test the significance of a potential punishment reputation.

2. METHODS

A total of 163 students played in groups of seven to nine (after written informed consent was obtained). Players in isolated cubicles could push buttons inside a box that was connected to a switchboard by a tangle of cables (Wedekind and Braithwaite 2002). The experimenter read the game instructions (supplementary material) and distributed player IDs in a procedure that ensured full anonymity (Wedekind and Braithwaite 2002). Each player received an initial amount of 20 CHF that could be used in a game. Final gains were paid out as in Wedekind and Braithwaite (2002).

One player was put in the Donor role, another in the Receiver role. The Donor (indicated to the player via a small bulbs inside the box) could either refuse to donate or donate to the Receiver by paying 1 CHF for the other to receive 2 CHF (we donated the difference). The Receiver could then decide whether or not to "make the Donor lose money" by accepting a cost of 1 CHF for the Donor to lose 2 CHF. Then a next pair of players was chosen. Each player played once per round as Donor and twice per two rounds as Receiver (e.g. once in rounds 1 and 2 each, or twice in round 2 only). The total number of rounds (16) was not communicated to the players.

During the first 8 rounds, the Donor's decision (but not his/her ID in order to control for potential confounding reputational effects (Wedekind and Milinski 2000; Wedekind and Braithwaite 2002)) was displayed on a projector screen to all players. We also displayed the Receivers' ID and his/her punishment score, i.e. an arrow wandering on a scale from -5 or +5, starting at 0, changing +1 for every punished and -1 for every non-punished defection. From round 9 on, we either removed the punishment option but continued to display the last punishment reputation the players had earned ("NOPUN/REP", n=36, 4 groups), left the option to punish but removed the display of reputation ("PUN/NOREP", n=36, 4 groups), or changed nothing ("PUN/REP", n=91, 11 groups; the larger number of groups under these control conditions allowed us to further analyze the within-group correlations between net

gain and punishment reputation after 16 rounds of undisturbed interactions). The "NOPUN/REP" treatment allowed us to test whether donors would either reward punishers for incurring the cost of disciplining non-cooperators or just stop discriminating between punishers and non-punishers in the absence of the threat of punishment. The "PUN/NOREP" treatment allowed us to assess the effect of punishment alone on the cooperation frequency. Players in the later two treatments were only informed at the end of round 8 about the change of rules.

The statistical analyses were carried out with R 2.10.1 (R Development Core Team 2010). We used the lme4 package for linear and logistic mixed-effect models (Bates and Sarkar 2007) that it is suitable for unbalanced designs (Baayen et al. 2008). Linear mixed-effect models were used to analyze group cooperation and punishment frequency during rounds 1 to 8 and 9 to 16, with group as random effect. Generalized linear mixed-effect models were used to analyze Donors' probability of giving, with group and donor as random effects.

3. RESULTS

By the end of the first 8 rounds, high probabilities of punishing defection lead to high levels of cooperation (Fig. 1a), and more cooperative groups reached higher total payoffs than less cooperative groups ($r = 0.74$, $n = 19$, $P < 0.001$). Within the controls ("PUN/REP") the effects of reputation did not seem to change over the full 16 rounds (likelihood ratio test (LRT), interaction between punishment score and round: $\chi^2 = 0.12$, $P = 0.73$; effect of round: $\chi^2 = 2.24$, $P = 0.13$). The within-group correlation between net gains (i.e. sum of received donations - punishment costs) and final punishment reputation ranged from $r = -0.80$ to 0.78 in the controls. These within-group correlations could be significantly predicted by a discriminant score that was the mean difference between the Receivers' punishment score when the Donor gave and did not give (Fig. 1b; the early interactions during a game seemed important here: the more individuals punished during the first 4 rounds, the higher their discrimination score at the end of the game, linear mixed-effect model, $t = 2.23$, $P = 0.03$).

During the experimental stage, i.e. from round 9 to 16, the frequency of cooperation did not significantly change in the controls (Fig. 2a; LRT: $\chi^2 = 2.55$, $P = 0.11$) but declined in both the PUN/NOREP and NOPUN/REP treatments (Fig. 2a; Table 1). In parallel, the probability of punishing defection declined when punishment could no longer affect reputation (Fig. 2b; Table 1). In the controls, Donors were more likely to give to Receivers with high punishment score than to those with a low punishment score (Fig. 2c, Table 1). In the PUN/NOREP treatment where no further punishment reputation could be built up, the updated punishment score that would correspond to the Receivers' actions but was no more displayed did also not seem to affect the Donors' decisions (LRT: $\chi^2 = 0.25$, $P = 0.61$; Fig. 2c). Correspondingly, the within-group correlation between the players' account and their probability of punishing defection was significantly higher in the controls than in the PUN/NOREP treatment (Welch t-test: $t_{12.95} = 3.14$, $P = 0.007$). In the NOPUN/REP treatment, where punishment was no longer an option, the reputation that had been built up until round 8 did not seem to affect the Donors' decisions either (LRT: $\chi^2 = 0.09$, $P = 0.76$; Fig. 2c).

4. DISCUSSION

Recent theory predicts that punishment has either evolved in another context than cooperation (Dreber et al. 2008; Wu et al. 2009) or that reputational effects compensate for the costs of punishment (Hilbe and Sigmund 2010; dos Santos et al. 2011). We found the latter to be true. High levels of cooperation were maintained when punishment could build up a punishment reputation, and, on average, the increased cooperation fully compensated for the costs of punishment. In groups with a high degree of discrimination between punishers and non-

punishers, the additional cooperation that punishers received even lead to net benefits, i.e. the costs of punishment were then overcompensated.

Punishers in public goods games may often be perceived as trustworthy and group focused, and may enjoy similar reputational benefits than generous people do in simple helping games (Barclay 2006). Human punishment has even been called "altruistic" (Yamagishi 1986; Fehr and Gächter 2000; Fehr and Gächter 2002) because people may punish defectors even in anonymous one-shot interactions where no benefit could be gained. However, anonymous one-shot interactions were probably very rare in human history, i.e. punishment is unlikely to have been evolved in such interactions (dos Santos and Wedekind 2012). In our game where players could built up a reputation in repeated interactions, punishers seemed feared rather than rewarded for altruistic behavior: Donors stopped discriminating between punishers and non-punishers when the opportunity to punish had been removed, i.e. the punishment reputation that had been built up before had no more effect when the threat of punishment was removed. As a consequence, those who had invested into their punishment reputation could not get compensated during the second part of our experiment and finally finished with relatively low payoffs.

When humans can observe the others' actions and can chose with whom to interact, they sometimes seem to weight cooperation higher than punishment (Rockenbach and Milinski 2011). Punishment turned out to be important in our experiments, but participants could not choose their partners and it is possible that the likelihood of punishing defection depends on whether partner choice is allowed for. Humans also tend to punish more often as the number of observers increases (Kurzban et al. 2007). This suggests that punishment is a strategic decision that takes aspects of the social environment into account. We observed that the outcome of the first interactions within a newly built social group influenced later dynamics: the participants' willingness to give to punishers depends on what they experienced at the beginning of a social interaction.

ACKNOWLEDGEMENTS

We thank the students for participation, R. Bergmüller, P. Bize, R. Bshary, M. Chapuisat, C. Clavien, C. El Mouden, A. Gardner, C. Hauert, M. Hochberg, L. Keller, C. Metzger, S. Nusslé, A. Ross-Gillespie, and S. West for comments or discussion, and the Swiss NSF for funding.

REFERENCES

Baayen, R. H., D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59:390-412.

Barclay, P. 2006. Reputational benefits for altruistic punishment. Evol. Hum. Behav. 27:325-344.

Bates, D. M., and D. Sarkar. 2007. lme4: Linear mixed-effects models using S4 classes, R package version 0.99875-9.

Brandt, H., C. Hauert, and K. Sigmund. 2003. Punishment and reputation in spatial public goods games. Proc. R. Soc. B 270:1099-1104.

Bshary, R., and A. S. Grutter. 2005. Punishment and partner switching cause cooperative behaviour in a cleaning mutualism. Biol. Lett. 1:396-399.

dos Santos, M., D. J. Rankin, and C. Wedekind. 2011. The evolution of punishment through reputation. Proc. R. Soc. B 278:371-377.

dos Santos, M., and C. Wedekind. 2012. Examining punishment at different explanatory levels. Behav. Brain Sci. 35.

Dreber, A., D. G. Rand, D. Fudenberg, and M. A. Nowak. 2008. Winners don't punish. Nature 452:348-351.

Fehr, E., and U. Fischbacher. 2003. The nature of human altruism. Nature 425:785-791.
Fehr, E., and S. Gächter. 2000. Cooperation and punishment in public goods experiments. Am. Econ. Rev. 90:980-994.
Fehr, E., and S. Gächter. 2002. Altruistic punishment in humans. Nature 415:137-140.
Gächter, S., E. Renner, and M. Sefton. 2008. The long-run benefits of punishment. Science 322:1510-1510.
Gardner, A., and S. A. West. 2004. Cooperation and punishment, especially in humans. Am. Nat. 164:753-764.
Hilbe, C., and K. Sigmund. 2010. Incentives and opportunism: from the carrot to the stick. Proc. R. Soc. B 277:2427-2433.
Hilbe, C., and A. Traulsen. 2012. Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. Sci Rep 2.
Kurzban, R., P. DeScioli, and E. O'Brien. 2007. Audience effects on moralistic punishment. Evol. Hum. Behav. 28:75-84.
Milinski, M., D. Semmann, and H. J. Krambeck. 2002. Reputation helps solve the 'tragedy of the commons'. Nature 415:424-426.
Nowak, M. A., and K. Sigmund. 1998. Evolution of indirect reciprocity by image scoring. Nature 393:573-577.
Nowak, M. A., and K. Sigmund. 2005. Evolution of indirect reciprocity. Nature 437:1291-1298.
R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Raihani, N. J., A. Thornton, and R. Bshary. 2012. Punishment and cooperation in nature. Trends. Ecol. Evol. 27:288-295.
Rand, D. G., A. Dreber, T. Ellingsen, D. Fudenberg, and M. A. Nowak. 2009. Positive interactions promote public cooperation. Science 325:1272-1275.
Rankin, D. J., M. dos Santos, and C. Wedekind. 2009. The evolutionary significance of costly punishment is still to be demonstrated. Proc. Natl Acad. Sci. USA 106:E135-E135.
Rockenbach, B., and M. Milinski. 2006. The efficient interaction of indirect reciprocity and costly punishment. Nature 444:718-723.
Rockenbach, B., and M. Milinski. 2011. To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. Proc. Natl Acad. Sci. USA 108:18307-18312.
Sigmund, K. 2007. Punish or perish? Retaliation and collaboration among humans. Trends Ecol. Evol. 22:593-600.
Tao, Y., C. Li, J. J. Wu, and R. Cressman. 2009. Reply to Rankin et al.: The efficiency ratio of costly punishment. Proc. Natl Acad. Sci. USA 106:E136-E136.
Ule, A., A. Schram, A. Riedl, and T. N. Cason. 2009. Indirect punishment and generosity toward strangers. Science 326:1701-1704.
Wedekind, C., and V. A. Braithwaite. 2002. The long-term benefits of human generosity in indirect reciprocity. Curr. Biol. 12:1012-1015.
Wedekind, C., and M. Milinski. 2000. Cooperation through image scoring in humans. Science 288:850-852.
Wu, J. J., B. Y. Zhang, Z. X. Zhou, Q. Q. He, X. D. Zheng, R. Cressman, and Y. Tao. 2009. Costly punishment does not always increase cooperation. Proc. Natl Acad. Sci. USA 106:17448-17451.
Yamagishi, T. 1986. The provision of a sanctioning system as a public good. J. Pers. Soc. Psychol. 51:110-116.

Figure 1. Cooperation and net gains when punishment reputation was unconstrained. (a) Frequency of cooperation and of punishing defection per group at the end of the first 8 rounds (Pearson's r = 0.54, n = 19 groups, P = 0.015). (b) Relationship between the within-group discriminant score (= the mean difference between the Receivers' punishment score when the Donor gave and did not give) and the correlation between net gains (i.e. sum of received donations - punishment costs) and punishment score in the PUN/REP treatment (Spearman's rank order correlation coefficient = 0.84, n = 11, P = 0.002).

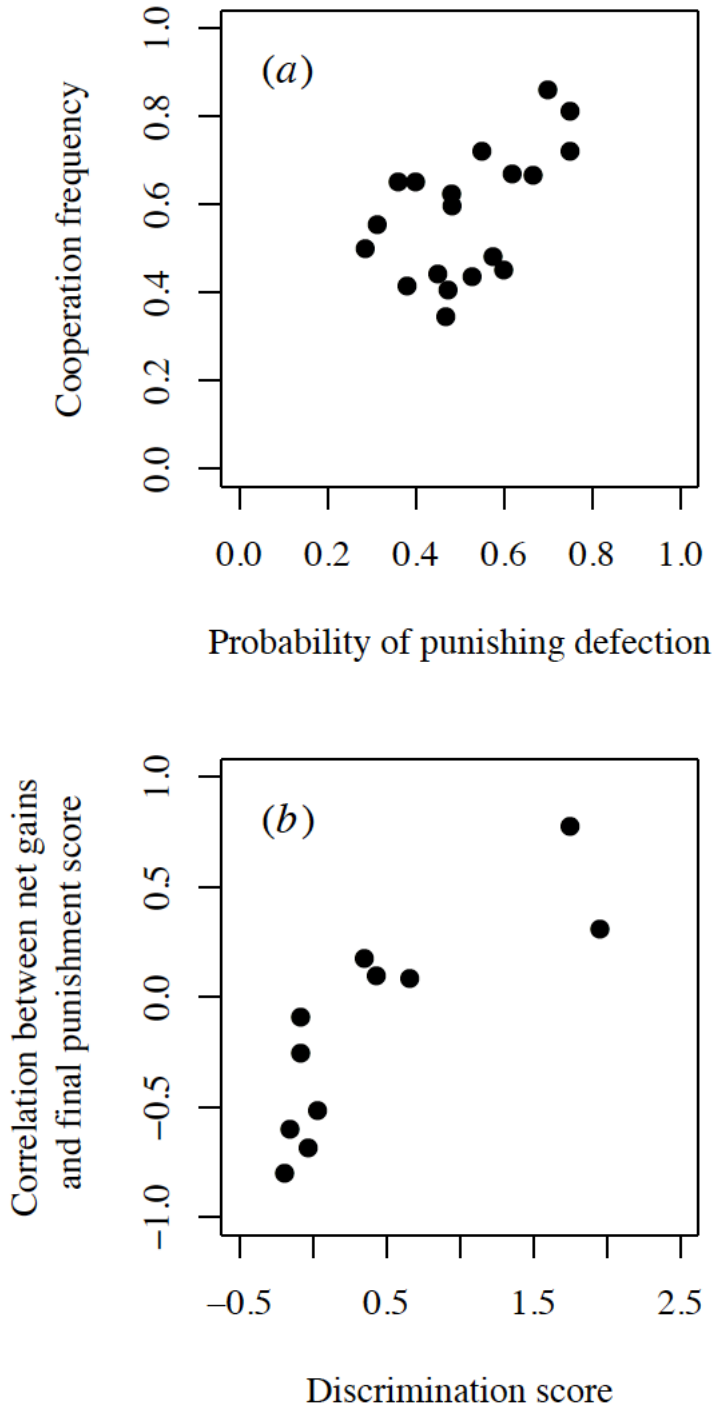

Figure 2. Treatment effects on (a) cooperation frequency, (b) the probability of punishing defection, and (c) the Donors' probability of giving to Receivers' with high or low punishment scores (corrected for group and round effects). PUN/REP = punishment was always possible and punishment reputation was continuously updated and displayed; PUN/NOREP = punishment reputation was no longer displayed in Part 2; NOPUN/REP = punishment was no longer possible in Part 2, but the reputation that had been built up during Part 1 was displayed. See Table 1 for statistics.

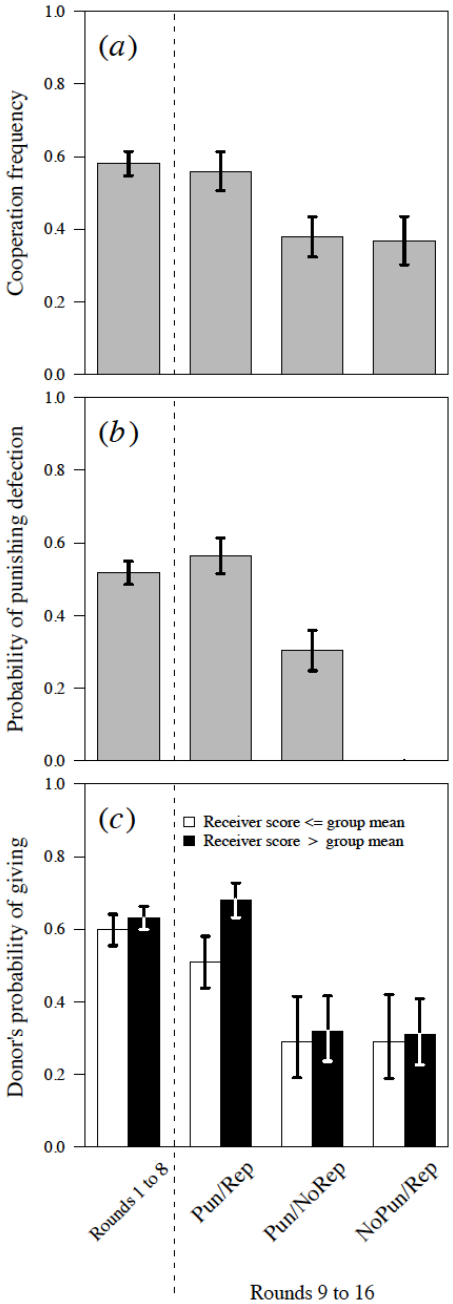

Table 1. The effects of experimental treatment and the punishment score that built up during the second part of the experiment (i.e. rounds 9-16) on (a) the frequency of cooperation, (b) the probability of punishing defection, and (c) the Donors' probability of giving. Linear (a and b) and generalized mixed-effects models (c) were fitted with and without a given effect (or interaction) in order to test if the goodness of fit between both models differed in a likelihood ratio test.

	χ^2	d.f.	P
a) Cooperation frequency			
Treatment	6.1	2	0.048
b) Probability of punishing defection			
Treatment	4.9	1	0.027
c) Donor's probability of giving			
Punishment score	8.2	1	0.004
Treatment	6.8	2	0.033
Punishment score x Treatment	2.9	2	0.235