Egalitarian Punishment in Humans

James Fowler[1], Tim Johnson[2], Richard McElreath[3], & Oleg Smirnov[4]

[1]*Department of Political Science, University of California, Davis 95616, USA.*

[2]*Max Planck Institute for Human Development, Berlin 14195, Germany.*

[3]*Department of Anthropology, University of California, Davis 95616, USA.*

[4]*Department of Political Science, University of Miami, Miami 33124, USA.*

Participants in laboratory public goods games are often willing to decrease the earnings of others at a cost to themselves.[1] What motivates this behaviour is unclear: in the conventional public goods game, punishment aimed at promoting cooperation cannot be distinguished from punishment seeking to attain equality.[2] To resolve this problem and better understand punishment behaviour, we deviate from the public goods framework and create an experimental game that isolates egalitarian motives. Subjects are placed in small groups and each is allocated a randomly determined sum of money; subjects are shown the amount given to other group members and are allowed to reduce others' incomes at a personal cost. Results show that individuals punish, at a cost to themselves, even when no cooperation norm can be enforced. The magnitude of punishment increases with income and, furthermore, the group member with the most income receives a disproportionate amount of punishment. Costly punishment of top earners suggests that relative status concerns influence human social competition. When employed in cooperative games where income correlates positively with defection, such behaviour could be misperceived as the sanctioning of defection.

Punishment in human and non-human animal societies generates selective pressure for various behaviours, most notably altruism and cooperation.[3-6] In particular,

altruistic punishment – the act of decreasing the earnings of others at a cost to oneself – has been shown to inspire and sustain cooperative behaviour in public goods games.[1,7,8] Although the act of punishment – and its ability to encourage cooperative behaviour – is unambiguous in such games, the underlying motivation for punishment is not. Since a player's contribution to the public good is proportional to her payoff from the public good, decreasing the payoff of a defector also has the effect of retrieving social equality. It therefore remains unclear whether low contribution or high payoff inspires punishment.[2]

Both motivations find indirect support in behavioural game research. Punishment of low contribution is suggested in studies indicating that good intentions generate trustworthiness[9] and ill intentions produce spiteful behaviour.[10] Punishment of high payoffs is suggested in research maintaining that egalitarian preferences inspire subjects to sacrifice individual benefit in order to obtain a uniform distribution of wealth.[11-13] Although these studies offer promising hypotheses, they do not clarify what triggers punishment. Direct studies of punishment offer no greater insight, as they yield unclear results[14] or fail to separate intentions from outcomes.[15,16] To avoid these problems, we use a simple experimental design to isolate the egalitarian punishment motivation.

The game we design closely resembles a widely-cited public goods game with punishment (see Methods).[1] Subjects are divided into groups of four *anonymous* members each. Subjects do not, however, make contribution decisions and there is no public goods aspect to the game. Instead, each player receives a sum of money randomly generated by a computer. Subjects are shown the *payoffs* of other group members and are then given an opportunity to give "negative tokens" to other players. Each negative token reduces the purchaser's payoff by 1 monetary unit (MU) and decreases the payoff of a targeted individual by 3 MUs. Groups are randomized after each round to prevent reputation from influencing decisions. We ensured that all

interactions between players were strictly anonymous**;** players were made aware that it is not possible to discover from whom one received negative tokens or to whom negative tokens were given.

Egalitarian punishment

Over the five sessions punishment was frequent: 62% of participants punished at least once, 31% punished five times or more, and 9% punished ten times or more. Most (67%) punishments were targeted at above-average earners in each group, with the top earner receiving about 44% of the total. The total amount of punishment increased with the income of the target (Fig. 1). Subjects who earned 35MUs or more received an average punishment of 7.9MUs compared to 0.5MUs for those who earned less than 21MUs. Income rank also appears to be important (Fig. 1). The top earner in each group received a disproportionate amount of punishment (7.0MUs) compared to other players (2.8MUs).

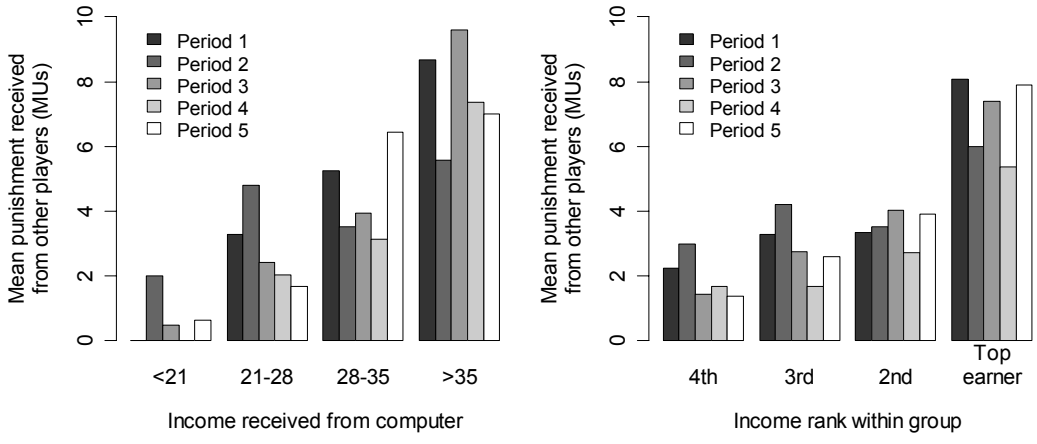

Figure 1 Mean punishment received in each period as a function of the income randomly allocated to the punished group member by the computer (left) and their income rank within the group (right).

Since punishment is costly and yields no material gain, self-interested subjects have no incentive to engage in it. As a result, we might expect punishments in our experiment to decline over time as subjects learn not to punish. Period-specific punishments (Fig. 1) show no consistent pattern over time, however. The mean punishment received in period 5 (4.08 MU) is actually *higher* than the punishment received in periods 1-4 (3.98 MU); a Wilcoxon signed rank test rejects the hypothesis that punishments are declining ($p<0.0001$). Initial mistakes do not explain punishment – subjects continue to punish after acquiring experience playing the game.

To explore whether absolute and relative income have independent effects on punishment behaviour, we conduct a Tobit regression of punishment received as a function of one's income controlling for being the top earner in the group. This method accounts for the censored nature of the punishment variable and the possibility that errors are correlated within each group. The regression coefficient on 'income' is 0.47 ($z=3.51$, $p=0.0005$), suggesting that subjects payoffs are reduced by about half an MU for each additional MU of income they receive from the computer. Above and beyond penalizing high incomes, top earners appear to be especially vulnerable to punishment. The regression coefficient on a 'top earner' dummy variable indicates that the person with the most income in each group received an additional 4.28 MUs in punishment ($z=3.51$, $p=0.0005$). Including variables for other income ranks does not improve the model; when added to the regression model as an explanatory variable, a dummy variable for the subject ranked 4^{th} in the group is insignificant ($z=-1.09$, $p=0.27$). To examine whether subjects might punish others based on their relative performance within the group, we added a variable that expressed income as a positive or negative deviation from the group average. This 'income deviance' variable is insignificant ($z=0.44$, $p=0.66$), suggesting that subjects focus on relative performance only to the extent that they punish the top earner. A stronger test of this hypothesis using a general additive model (GAM) with poisson distribution and estimating a cubic spline for both

'income' and 'income deviance' ensured the top earner result was not due to misspecification of the effect of absolute and relative income on punishment. The coefficient on 'top earner' in the GAM regression was strongly significant (t=5.75, p<0.0001).

It is important to emphasize that there is no material benefit to punishment in this game. Income is drawn from a random distribution independent of subjects' decisions; no amount of punishment can increase future payoffs for anyone. Subjects were aware of this fact. Moreover, a desire for revenge cannot explain the punishment. Subjects are told that they never meet the same person twice, so they cannot satisfy a desire to reciprocate any punishment they have received in future rounds.

Do emotions cause egalitarian punishment?

Others[1] show that experimental subjects feel anger towards free-riders in a public goods setting and claim that this anger may motivate the punishment of non-contributors. In our experiment there are neither contributions nor a public good, so we wondered why people engage in punishment. One possibility is that inequality arouses negative emotions that instigate punishment. If so, we should observe annoyance and anger at high earners, and these sentiments should increase as inequality increases.

To elicit emotional reactions, we presented subjects with hypothetical scenarios in which they encountered group members who obtained higher payoffs than they did (see methods). Subjects were then asked to indicate on a 7 point scale whether they felt annoyed or angry (1='not at all', 7='very much') by the other individual. In one scenario, subjects were told they encountered an individual whose payoff was considerably greater than their own. This scenario generated much annoyance: 68% of the subjects claimed to be at least somewhat annoyed, while 32% indicated a high level

(4 or more) of annoyance. Many subjects (42%) also indicated they felt some anger towards the top earner. In another scenario inequality between subjects' incomes was smaller, and there was significantly less anger (Wilcoxon signed rank test, p=0.003) and annoyance (p<0.0001). Only 41% indicated they were annoyed and 26% indicated they were angry. Individuals apparently feel negative emotions toward high earners and the intensity of these emotions varies with income inequality.

We were also interested in the top earner's *expectations* of others' emotions, so we presented subjects with other hypothetical scenarios in which they were the top earner. Subjects were asked to rate how angry they believed others would be if they encountered them. The level of inequality was varied in the same manner as in the first hypothetical scenarios (see methods). We found that when inequality was considerable subjects *expected* significantly more anger (Wilcoxon signed rank test, p=0.0003) and annoyance (p=0.009) than they themselves expressed. Most subjects (80%) expected others to be at least somewhat annoyed and 44% of them expected the level of annoyance to be high (4 or more); many (67%) expected others to feel anger towards them. When the difference in income was reduced, it generated significantly less anger (Wilcoxon signed rank test, p<0.0001) and annoyance (p<0.0001). Only 56% indicated that they expected others to be annoyed and 45% expected others to be angry.

People appear to feel negative emotions towards top earners and most expect others to be angry or annoyed when they themselves receive the most income. These emotions influence punishment decisions. Subjects who said they were at least somewhat annoyed or angry at the top earner in the hypothetical scenarios spent significantly more on punishment in the experiment than those who said they were not (Fig. 2). A Tobit regression shows that the magnitude of anger (on the 7 point scale) also significantly increases spending on punishment (coefficient 0.24, z=2.54, p=0.01).

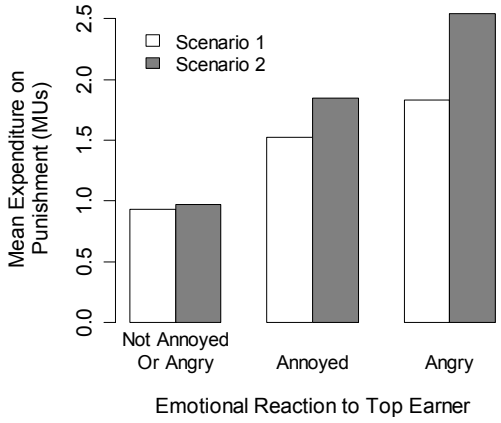

Figure 2. Mean expenditure on punishment in each period as a function of feelings towards the top earner in two hypothetical scenarios (see Methods). The top earner receives more relative to the group in scenario 1 than in scenario 2. Subjects who indicated they were at least somewhat annoyed or angry in the hypothetical scenarios punish significantly more in the experiment than those who do not (two-sided *t*-test, $p<.003$ for all comparisons). Subjects who expressed the greatest sensitivity to inequality in the hypothetical scenarios (those who were annoyed or angered at the top earner even when the income difference was low in scenario 2) tended to deliver the most punishment in the experiment.

The results indicate that social inequality arouses negative emotions that motivate punishment. This finding supports research that suggests humans are motivated by egalitarian preferences.[11-13] It is also consistent with the punishment of non-contributors in public good games.[1] Egalitarian motives, in addition to direct motives to punish free-riders, may therefore underlie the sanctioning of free-riders and facilitate the maintenance of cooperation.

Finally, standard evolutionary models suggest that strategies that attend to relative payoffs within local groups will not evolve, unless populations are regulated locally.[17-19] Human populations tend to be well-mixed, as evidenced by measured amounts of genetic variation among local groups[20], implying populations are not strongly regulated locally. This makes puzzling the observation of deliberate costly punishment aimed at changing relative payoffs within the group. Future models will need to reconcile the experimental results presented here with existing theory.

Methods

The design and procedures of the experiment closely approximate a widely-cited public good experiment.[1] One hundred (n=100) students from the University of California at Davis volunteered to participate in the experiment. Recruitment of subjects was conducted in several different departments to maximize the chance that subjects did not know one another; any student who was at least 18 years old was eligible to take part in the study. Twenty subjects attended each of the five experimental sessions and each session involved five periods. Every period, subjects were randomly placed in groups of four subjects. At the beginning of each period subjects received a random payoff and were shown the payoffs for all four members of their group. To maintain comparability with other public goods games, random payoffs were drawn from the empirical distribution of payoffs in the first stage of a widely-cited public goods game with punishment.[1] Subjects were then given an opportunity to punish any member of the group by purchasing up to 10 negative tokens for each player. At the end of each period, subjects learned the amount of punishment they received and their new payoff. The experiment lasted 30 minutes and on average subjects earned approximately 10 dollars per session.

All activity in the experiment was completely anonymous. Group composition changed every period so that no one played with the same person more than once. The

subjects were ignorant of other players' experimental history: neither past payoffs nor past punishment decisions were known. Different group composition each period and the absence of any history of play ensured that subjects could neither develop reputations nor target other subjects for revenge.

At the beginning of each session subjects were asked to read experiment instructions on their individual computer screens (available from the authors on request), and they also had a paper copy available for reference. The instructions explained all features of the experiment, including how payoffs are determined, how group composition is altered every period, and how anonymity of individual decisions and payoffs in the experiment is preserved. In order for the experiment to start, subjects had to answer *correctly* several test questions designed to ensure full understanding of how choices in the game generate payoffs. All interactions between players were made via computer terminals using the experimental software GameWeb.[21] At the end of the experimental session, subjects were asked to complete a survey about their demographic characteristics and a questionnaire concerning emotions.

The emotions questionnaire presented four hypothetical scenarios to subjects. The first two were "You receive 23 [19] tokens. The second group member receives 25 [21] and the third 21 [17] tokens. Suppose the fourth member receives 37 [22] tokens. You now accidentally meet this member. Please indicate your feelings towards this person." (Unbracketed numbers were used in the first scenario and bracketed numbers were used in the second scenario.) The second two scenarios were: "Imagine that the other three group members receive 21 [17], 23 [19], and 25 [21] tokens. You receive 37 [22] tokens and the others know this. You know accidentally meet one of the other members. Please indicate the feelings you expect from this member towards you." After reading each scenario, subjects were asked to indicate on a 7 point scale whether they felt annoyed or angry (1='not at all', 7='very much').

The authors declare that they have no competing financial interests.

We would like to thank the UC Davis Institute of Government Affairs for generous research support and … for helpful comments.

Correspondence and requests for materials should be addressed to J.F. (jhfowler@ucdavis.edu).

References

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

2. Fowler, J. H., Johnson, T. & Smirnov, O. Egalitarian motive and altruistic punishment. *Nature* 433, doi:10.1038/nature03256 (2005).

3. Boyd, R. & Richerson, P. J. Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups. *Ethology and Sociobiology* 13, 171-195 (1992).

4. Clutton-Brock, T. H. & Parker, G. A. Punishment in Animal Societies. *Nature* 373, 209-216 (1995).

5. Fowler, J. H. Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences of the United States of America* 102, 7047-7049 (2005).

6. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America* 100, 3531-3535 (2003).

7. Ostrom, E., Gardner, R. & Walker, J. *Rules, Games, and Common-Pool Resources* (University of Michigan, Ann Arbor, 1995).

8. Yamagishi, T. The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51, 110-116 (1986).

9. McCabe, K. A., Rigdon, M. L. & Smith, V. L. Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization* 52, 267-275 (1998).

10. Blount, S. When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes* 63, 131-144 (1995).

11. Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114, 817-868 (1999).

12. Boehm, C. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior* (Harvard University Press, Cambridge, 1999).

13. Bolton, G. & Ockenfels, A. ERC: A theory of Equity, Reciprocity, and Competition. *American Economic Review* 90, 166-193 (2000).

14. Guth, W., Kliemt, H. & Ockenfels, A. Retributive responses. *Journal of Conflict Resolution* 4, 453-469 (2001).

15. Falk, A., Fehr, E. & Fischbacher, U. On the nature of fair behavior. *Economic Inquiry* 41, 20-26 (2003).

16. Fehr, E. & Gachter, S. Egalitarian motive and altruistic punishment - Reply. *Nature* 433, doi:10.1038/nature03256 (2005).

17. Hamilton, W. D. Selfish and spiteful behaviour in an evolutionary model. *Nature* 228, 1218-1220 (1970).

18. Boyd, R. Density-dependent mortality and the evolution of social interactions. *Animal Behavior* 30, 972-982 (1982).

19. Taylor, P. D. Altruism in viscous populations--an inclusive fitness approach. *Evolutionary Ecology* 6, 352-356 (1992).

20. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* (Sinauer Associates, Sunderland, MA, 1997).

21. McElreath, R. (Davis, CA, 2005).