Inequality Reduces Punishment-Induced Cooperation 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.*

Humans often cooperate, voluntarily paying an individual cost to supply a benefit to others. Public good experiments show that punishment induces a high level of cooperation, even when it is costly to the punisher.[1] It is unclear, however, what motivates individuals to engage in costly punishment: a desire to retaliate against non-cooperators or a desire to reduce inequality among group members. Although both motives might have a positive effect on cooperation[2-4], they cannot be separated in the conventional public good game.[5] Here we conduct an experiment in which we add a randomly-generated fixed payoff to a public good game with punishment. This design allows us to determine whether punishment is aimed at low contributors or high earners. The results show that players punish frequently, penalizing *both* those who contribute the least and those who earn the most. However, the exogenously-created inequality tends to distort the meaning of punishment, which dramatically reduces the amount of cooperation observed. This evidence suggests that social equality may be necessary if punishment is to have a positive influence on cooperation in humans.

In the public good game with punishment, individuals are endowed with a resource that can be contributed to a common pool. If contributed, the resource

increases in value and is divided equally among group members. Social welfare is maximized in the game if all group members contribute to the common pool, whereas personal wealth is greatest when an individual retains her endowment and others contribute fully. Given these conditions, punishment offers the opportunity to both penalize non-contribution and correct the inequality that non-contribution produces.[5] Exercising the punishment option sends a clear message to non-contributors: the benefits of non-contribution will be reduced, sometimes to the extent that contribution would be a more profitable strategy.

Here we study what happens when the possible motivations for punishment are no longer confounded and the clear message provided by punishment is distorted. Subjects are divided into groups of four and play a five period public good game. After each period group membership is altered such that no individual plays with another individual twice. At the beginning of each round, players make contribution decisions. Individual payoffs from the public good are then calculated and a randomly determined positive sum is added to this value. Subjects are then informed of their own payoff, as well as the payoffs and contributions of other group members. Subsequently, players are allowed to reduce the payoffs of others. Once punishment is completed, players learn their post-punishment payoffs for that round and the next round of the experiment begins.

Hypotheses based on previous research offer conflicting predictions about what to expect in the experiment. Existing evidence indicates that sanctions believed to serve selfish motives degrade trustworthiness,[6] suggesting that punishment motives unrelated to contribution enforcement might weaken the mechanisms that foster cooperation in public good games. Conversely, previous studies have shown that the enforcement of cooperative norms[2,3] and the maintenance of social equality[4] can facilitate cooperation independently, suggesting that punishment based on either motive might increase

cooperation. To determine whether punishment increases social cooperation when exogenous inequality is introduced and punishment motives are decoupled, we consider our experimental evidence.

Altruistic and egalitarian punishment

Punishment in the experiment was common: 69% of participants punished at least once, 45% punished five times or more, and 14% punished ten times or more. Most (75%) punishments were targeted at individuals who contributed less (75%) or earned more (73%) than the group average. The top contributor in each group received about 36% of the total punishments and the top earner received about 40%. The average size of the punishments increased with the difference between the average group contribution and the contribution by the target (Fig. 1). For example, subjects spent an average 4.9MUs to punish those whose contribution fell short of the group average by 8 MUs or more, compared to 1.4MUs for those who contributed about the same as other group members (+/-3 MUs). Income rank also appears to be important (Fig. 1). Subjects spent a disproportionate amount on the punishment of top earners (3.9MUs) compared to other players (1.8MUs).

Since punishment is costly and cannot influence the behaviour of future group members, 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 expenditure on punishment in period 5 (2.6 MU) is actually *higher* than the punishment received in periods 1-4 (2.3 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.

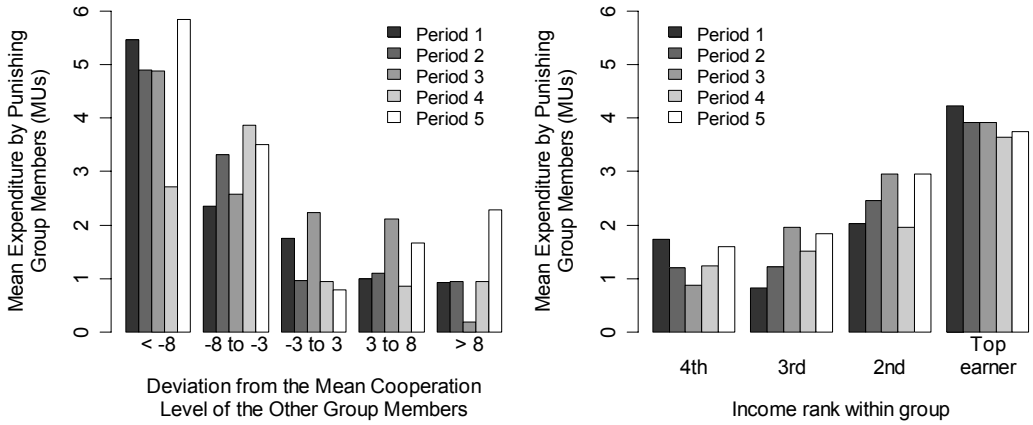

Figure 1 Mean expenditure on punishment in each period as a function of the deviation of the contribution of the punished group member from the mean contribution of the other members (left) and their income rank within the group prior to punishment (right).

To explore whether income and contribution behaviour have independent effects on punishment behaviour, we conduct a Tobit regression of punishment received. This method accounts for the censored nature of the punishment variable and the possibility that errors are correlated within each group. Independent variables in this regression are the positive and negative deviation of a member's contribution from the average group contribution and a dummy variable indicating whether the member has the highest income in the group. The regression coefficient on 'negative deviation' is 0.41 (z=6.81, p<0.0001), suggesting that subjects spend a little less than half an MU on punishment for each additional MU that the target's contribution falls short of the average group contribution. The coefficient on 'positive deviation' is -0.12 (z=-2.15, p=.03), indicating above-average contributors are punished somewhat less as their contributions increase.

Above and beyond the penalizing of low contributors, top earners appear to be especially vulnerable to punishment. The regression coefficient on 'top earner' indicates that the group spends an extra 1.34 MUs (z=2.67, p=0.008) to punish the

highest income member of the group. Including a dummy variable that indicates whether the member was the lowest contributor in the group does not substantially change the coefficient on 'top earner' (1.20, z=2.45, p=.01). Thus, high income earners are apparently targeted for punishment independently of their contribution behaviour. A stronger test of this hypothesis using a general additive model (GAM) with Poisson distribution and estimating a cubic spline for both 'contribution' and 'contribution deviance' ensured the top earner result was not due to misspecification of the effect of absolute and relative contribution on punishment. The coefficient on 'top earner' in the GAM regression was strongly significant (t=4.91, p<0.0001). Thus, punishment behaviour appears to be motivated by a mix of concerns for cooperation and equality.

Inequality and Cooperation

It has been shown that adding punishment to a public goods game significantly increases cooperation.[1] Many subjects in these experiments anticipate that low contributions will be punished, so they increase their contributions. Others learn to increase their contributions over time in response to punishment. As a result, contributions in games with punishment typically start at a medium-high level and converge to a higher level over time.[1-3]

In contrast, our experiment shows that adding a random payoff to a public goods game with punishment significantly reduces cooperation. The average contribution for all periods in our experiment is 7.4MUs (S.E. 0.3), and it remains stable over time (Fig. 2). Cooperation is significantly lower in every period than that reported in a widely-cited public good game with punishment[1] (Wilcoxon rank test, p<.0001). In fact, contributions in the final period of the game with additional random payoffs are less than half those in the game without. It appears that inequality reduces punishment-induced cooperation.

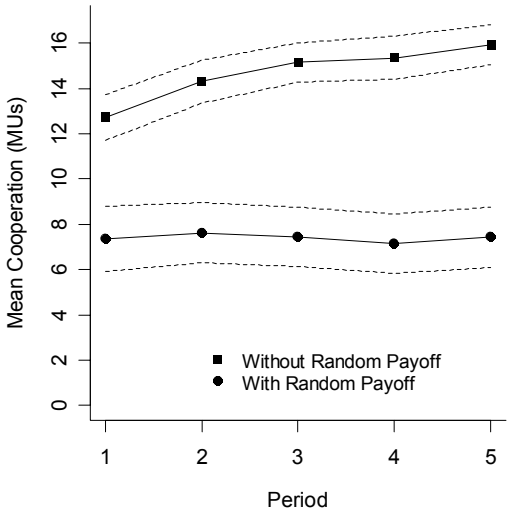

Figure 2. Time trend of mean cooperation together with the 95% confidence intervals from a widely-cited public goods game with punishment[1] (top line) and from a public goods game with punishment in which individuals were given additional random payoffs from the computer (bottom line).

When payoffs are unequal for reasons not related to contributions, subjects may have difficulty interpreting the punishment they receive—punishers might be attempting to restore equality rather than promote cooperation. As a result, people are less likely to respond to observed punishment by increasing contributions. More importantly, subjects appear to anticipate this problem by offering low contributions in the first round.

We test this hypothesis by examining the effect of punishment on the change in a subject's contribution from period t-1 to period t (Fig. 3). Subjects who were not punished *decreased* their contribution by 0.81MUs in the game with additional random income and by 0.69MUs in the game without. A two-sided t test suggests that reactions to the absence of punishment in these two treatments are not significantly different (p=0.75). In contrast, subjects who did receive punishment *increased* their contribution by 0.62MUs in the game with additional random income and by 1.47MUs in the game without. These results indicate that adding random income to the public goods game

significantly decreases the effect of punishment on future contributions (one-sided *t*-test, p=0.03).

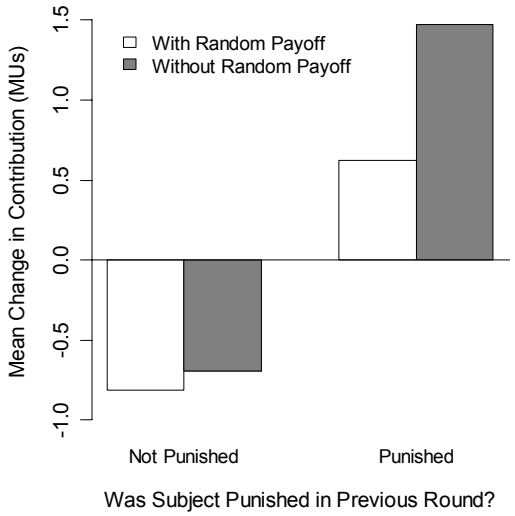

Figure 3. Effect of punishment on change in contributions in a widely-cited public goods game with punishment[1] (gray) and a public goods game with punishment in which individuals were given additional random payoffs (white). Subjects reacted significantly more strongly to punishment in the game *without* additional random payoffs.

It is important to emphasize that subjects had *exactly* the same information about income and contributions in each experiment when they were making punishment decisions. The only difference between experiments was the exogenously created inequality. Consistent with other experimental results[4] and social theory[7] this inequality appears to have eroded social cooperation.

Our experiment provides further evidence that the interpretation of punitive motives affects mechanisms that produce social cooperation.[6] It also supports results from formal models that indicate the evolution of cooperation relies on conformist pressures that reserve punishment for defectors.[8] Punishment provides a viable pathway

to cooperation only if penalties are directed toward non-contributors. More importantly, social equality might be an important condition for this pathway, since punishment in unequal environments has mixed meanings.

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 were endowed with 20MUs and permitted to contribute any portion of these towards a group project. Each MU invested in the group project yielded 0.4MUs for each group member. After the contribution stage, each player received a random payoff and were shown the contributions and total 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 retribution.

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.[9] At the end of the session, subjects were asked to complete a survey about their demographic characteristics.

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. Ostrom, E., Gardner, R. & Walker, J. *Rules, Games, and Common-Pool Resources* (University of Michigan, Ann Arbor, 1995).
3. Yamagishi, T. The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* **51**, 110-116 (1986).

4. Anderson, L.R., Mellor, J.M., and Milyo, J. Inequality and Public Good Provision: An Experimental Analysis. *College of William and Mary Department of Economics Working Paper*. (2004).
5. Fowler, J. H., Johnson, T. & Smirnov, O. Egalitarian motive and altruistic punishment. *Nature* **433**, doi:10.1038/nature03256 (2005).
6. Fehr, E. & Rockenbach, B. Detrimental effects of sanctions on human altruism. *Nature* **422**, 137-140 (2003).
7. Putnam, R. *Bowling Alone*. (Simon and Schuster, New York, 2000).
8. Henrich, J. & Boyd, R. Why people punish defectors. *Journal of Theoretical Biology* **208**, 79-89.
9. McElreath, R. (Davis, CA, 2005)