Human behaviour

Egalitarian motive and altruistic punishment

Arising from: E. Fehr & S. Gächter *Nature* **415,** 137–140 (2002)

Altruistic punishment is a behaviour in which individuals punish others at a cost to themselves in order to provide a public good. Fehr and Gächter[1] present experimental evidence in humans indicating that negative emotions towards non-cooperators motivate punishment, which, in turn, provokes a high degree of cooperation. Using Fehr and Gächter's original data, we provide an alternative analysis of their experiment that suggests that egalitarian motives are more important than motives for punishing non-cooperative behaviour. This finding is consistent with evidence that humans may have an evolutionary incentive to punish the highest earners in order to promote equality, rather than cooperation[2].

In the experiment by Fehr and Gächter, groups with four members played a public-good game. Each participant was given an initial endowment of 20 money units (MUs), which they could either keep or contribute (entirely or partially) to a group project. For every MU invested in the project, each member earned 0.4 MU. Although the dominant strategy in the game is to keep the whole endowment, mutual contribution yields the best result for the group. In one treatment, subjects had an option to decrease the payoff of other group members, such that 1 MU spent on punishment decreased the payoff of the targeted individual by 3 MUs. The punishment stage started immediately after subjects had seen the payoffs earned by other group members in the first stage.

Punishment in the experiment was frequent and followed a pattern. Most negative points were imposed on below-average contributors and those who earned above-average payoffs in the first round. Fehr and Gächter define defection in relative terms, asserting that subjects punish an individual j in proportion to his or her deviance from the mean contribution of the other three players:

$$\frac{1}{3}\sum_{i\neq j}(c_i) - c_j$$

However, suppose individuals were not concerned about contributions and instead wanted to minimize inequality in the payoffs. If so, they might choose punishments in proportion to payoff deviance:

$$\pi_j - \frac{1}{3}\sum_i \pi_i$$

Notice that, as

$$\pi_j = 0.4\sum_i (c_i) - c_j$$

in the Fehr and Gächter experiment, payoff deviance is exactly equal to contribution deviance:

$$0.4\sum_i (c_i) - c_j - \frac{1}{3}\sum_{i\neq j}\left(0.4\sum_i (c_i) - c_j\right) = \frac{1}{3}\sum_{i\neq j}(c_i) - c_j$$

Thus, it is not possible to tell them apart, and all of Fehr and Gächter's statistical results equally support the hypothesis that subjects are punishing the top earners in order to minimize the difference in payoff outcomes.

If absolute levels are used instead of deviance from the mean, the experiment suggests that payoffs are important in altruistic punishment. We replicated Fehr and Gächter's regression analysis of the data and then used the same method to examine how group expenditures for the punishment of player i varied with player i's contribution, prepunishment payoff, and an interaction between the two.

The resulting model suggests that the payoff has a strong and significant effect on punishment, even controlling for the contribution. For example, a 10-MU increase in the payoff yields 6.1 MU (±1.1 MU) of additional punishment when the contribution is 0, and 1.8 MU (±1.4 MU) when the contribution is 20. By contrast, the contribution has less effect on punishment and only decreases punishment when the payoff is sufficiently high. A 10-MU increase in the contribution yields a 3.6-MU (±1.4 MU) decrease in the total punishment when the payoff is 44 (the maximum observed value), but the contribution has no significant effect on punishment when the payoff is 13 MU (the minimum observed value). These results indicate that subjects were more motivated to punish high earners than low contributors, and that egalitarian motives may underlie altruistic punishment in humans.

James H. Fowler*, Tim Johnson†, Oleg Smirnov‡

**Department of Political Science, University of California, Davis, One Shields Avenue, Davis, California 95616, USA*
e-mail: jhfowler@ucdavis.edu
†Max Planck Institute for Human Development, Center for Adaptive Behaviour and Cognition, 14195 Berlin, Germany
‡Department of Political Science, University of Oregon, 1284 University of Oregon, Eugene, Oregon 97403, USA

doi:10.1038/nature03256

1. Fehr, E. & Gächter, S. *Nature* **415,** 137–140 (2002).
2. Boehm, C. *Hierarchy in the Forest: The Evolution of Egalitarian Behaviour* (Harvard Univ. Press, Cambridge, 1999).

Fehr and Gächter reply — Fowler *et al.* raise an important question[1]. They correctly argue that the desire to reduce inequality may motivate cooperators who altruistically punish free riders in our experiments[2]. Also, the evolutionary history of humans suggests that egalitarianism shaped many human cultures[3] and that egalitarian motives may, therefore, be a powerful force behind the punishment of free riders. In addition, recently developed proximate theories[4], which formalize the notion of inequality aversion, also suggest that egalitarian desires may be important. Fowler *et al.* contrast their egalitarianism hypothesis with our view that negative emotions against free riders drive punishment.

However, the two views are not necessarily incompatible: egalitarian sentiments may be the basis behind cooperators' negative emotions because free riding causes considerable inequalities. Moreover, the reanalysis of our original data by Fowler *et al.* can only raise (but not settle) the question of whether equality motives are important because a punishing cooperator in our experiments[2] inevitably reduces the inequality between himself and the punished free rider. Thus, it is not possible to isolate any other motive behind altruistic punishment based on these data because the equality motive can never be ruled out.

A plausible alternative to the egalitarian motive is that cooperative subjects may perceive free riding as a violation of the strong reciprocity norm[5–7]. Cooperators may feel exploited by the free riders because the latter did not reciprocate their cooperative choices. Retaliation motives drive altruistic punishment in this view.

The retaliation motive has been isolated in a public-good experiment (A. Falk, E. F. and U. Fischbacher, see www.iew.unizh.ch/wp/iewwp059.pdf) in which the potential impact of the equality motive was removed. This experiment was almost identical to our original[2], except that punishment did not change the income difference between the punished and the punishing subject. One money unit (MU) spent on punishment reduced the free rider's payoff by exactly this amount. Thus, if egalitarian motives are the sole driving force behind altruistic punishment, there should be no punishment in this experiment. However, punishment is frequently observed (Fig. 1).

This punishment pattern is very similar to that of the original experiment because those who cooperate predominantly punish the free riders. Overall, subjects punish other group members in the new experiments 211 times: 51 out of 87 subjects (59%) punish at least once, and 22% punish more than five times during the experiment, which consists of six rounds. There is a considerable amount of punishment in the new experiments, although the equality motive cannot be