Altruistic punishment in humans

Ernst Fehr* **& Simon Gächter†**

** University of Zürich, Institute for Empirical Research in Economics, Blümlisalpstrasse 10, CH-8006 Zürich, Switzerland*
† University of St Gallen, FEW-HSG, Varnbüelstrasse 14, CH-9000 St Gallen, Switzerland

Human cooperation is an evolutionary puzzle. Unlike other creatures, people frequently cooperate with genetically unrelated strangers, often in large groups, with people they will never meet again, and when reputation gains are small or absent. These patterns of cooperation cannot be explained by the nepotistic motives associated with the evolutionary theory of kin selection and the selfish motives associated with signalling theory or the theory of reciprocal altruism. Here we show experimentally that the altruistic punishment of defectors is a key motive for the explanation of cooperation. Altruistic punishment means that individuals punish, although the punishment is costly for them and yields no material gain. We show that cooperation flourishes if altruistic punishment is possible, and breaks down if it is ruled out. The evidence indicates that negative emotions towards defectors are the proximate mechanism behind altruistic punishment. These results suggest that future study of the evolution of human cooperation should include a strong focus on explaining altruistic punishment.

Throughout evolution, crucial human activities like hunting big game, sharing meat, conserving common property resources, and warfare constituted a public good. In situations like these, every member of the group benefits from the 'good', including those who did not pay any costs of providing the good. This raises the question of why people regularly participate in costly cooperative activities like warfare and big-game hunting[1–4]. Several theories have been proposed to explain the evolution of human cooperation. The theory of kin selection[5] focuses on cooperation among individuals that are genetically closely related, whereas theories of direct reciprocity[6–9] focus on the selfish incentives for cooperation in bilateral long-term interactions. The theories of indirect reciprocity[10–15] and costly signalling[16–18] show how cooperation in larger groups can emerge when the cooperators can build a reputation. Yet these theories do not readily explain why cooperation is frequent among genetically unrelated people, in non-repeated interactions, when gains from reputation are small or absent.

Punishment provides a solution to this problem. If those who free ride on the cooperation of others are punished, cooperation may pay[3,19–23]. Yet this 'solution' begs the question of who will bear the cost of punishing the free riders. Everybody in the group will be better off if free riding is deterred, but nobody has an incentive to punish the free riders. Thus, the punishment of free riders constitutes a second-order public good. The problem of second-order public goods can be solved if enough humans have a tendency for altruistic punishment, that is, if they are motivated to punish free riders even though it is costly and yields no material benefits for the punishers.

We examined the question of whether humans engage in altruistic punishment and how this inclination affects the ability of achieving and sustaining cooperation. A total of 240 students participated in a 'public goods' experiment with real monetary stakes and two treatment conditions: punishment and no punishment. In both conditions, groups with four members played the following public goods game. Each member received an endowment of 20 money units (MUs) and each one could contribute between 0 and 20 MUs to a group project. Subjects could keep the money that they did not contribute to the project. For every MU invested in the project, each of the four group members, that is, also those who invested little or nothing, earned 0.4 MUs. Thus, the investor's return from investing one additional MU in the project was 0.4 MUs, whereas the group return was 1.6 MUs. Because the cost of investing 1 MU in the project was exactly 1 MU, whereas the private return was only 0.4 MUs, it was always in the material self-interest of any subject to keep all MUs privately—irrespective of how much the other three subjects contributed. Yet, if all group members kept all MUs privately, each subject earned only 20 MUs, whereas if all of them invested their 20 MUs each subject would earn $0.4 \times 80 = 32$ MUs.

All the interactions in the experiment took place anonymously. Members were not informed of the identity of the others in the group. Subjects made their investment decisions simultaneously and, once the decisions were made, they were informed about the investments of the other group members. The only difference between the two conditions was that in the punishment condition, subjects could punish each of the other group members after they were informed about the others' investments. A punishment decision was implemented by assigning between 0 and 10 points to the punished member. Each point assigned cost the punished member 3 MUs and the punishing member 1 MU. All the punishment decisions were also made simultaneously.

Because we conjectured that the opportunity for punishing would have a larger impact if subjects could learn about the behaviour of other group members, we repeated the basic public goods game—with and without punishment opportunity, depending on the treatment—for six periods. To rule out that repetition created cooperation or punishment through direct reciprocity[6–9] or reputation[10–15], the group composition changed from period to period such that no subject ever met another subject more than once. Moreover, our design ruled out any kind of reputation formation (see Methods), so purely selfish subjects will never cooperate or punish others, because cooperation and punishment are costly and yield no pecuniary benefits. Therefore, the selfish motives associated with theories of indirect reciprocity[10–15] or costly signalling[16–18] cannot explain cooperation and punishment in this environment.

However, punishment may well benefit the future group members of a punished subject, if that subject responds to the punishment by raising investments in the following periods. In this sense, punishment is altruistic. In the presence of altruistic punishers, even purely selfish subjects have a reason to cooperate in the punishment treatment.

Altruistic punishment and cooperation

Altruistic punishment took place frequently. In the ten sessions, subjects punished other group members a total of 1,270 times; 84.3% of the subjects punished at least once, 34.3% punished more than five times during the six periods, and 9.3% punished even more than ten times. Punishment also followed a clear pattern. Most (74.2%) acts of punishment were imposed on defectors (that is, below-average contributors) and were executed by cooperators (that is, above-average contributors), and punishment of the defectors was harsh (Fig. 1). For example, if a subject invested

14–20 MUs less than the average investment of the other group members during periods 5 and 6, the total group expenditures for punishing this subject were almost 10 MUs. Moreover, the more a subject's investment fell short of the average investment of the other three group members, the more the subject was punished. The pattern and strength of punishment was also stable across time (Fig. 1). A Wilcoxon signed rank test of punishment in periods 1–4 versus periods 5 and 6, with 10 matched observations, yields $z = -1.07$, $P = 0.285$ (two-tailed). The same test for periods 1–5 versus period 6 yields $z = 0.178$, $P = 0.859$ (two-tailed).

We examined how the group expenditures for the punishment of member i varied with the positive and the negative deviation of member i's cooperation from the average cooperation of the others. Our examination is based on Tobit regressions. The regression coefficient on 'negative deviation' is 0.622 ($z = 18.1$, $P < 0.0005$) and the coefficient on 'positive deviation' is −0.149 ($z = -2.86$, $P < 0.004$). The hypothesis tests associated with the regression are based on robust standard errors that take into account that only the observations across sessions are independent and that punishment is a censored variable. The average cooperation of other group members, if added as an explanatory variable to the regression, is insignificant ($z = -0.99$, $P = 0.322$). This analysis indicates that an increase in the negative deviation of i from the others' average cooperation by 10 MUs increased the punishment expenditure of the others by 6.22 MUs (and, hence, the pay-off reduction imposed on i by 18.66 MUs), whereas an increase in the positive deviation of i by 10 MUs reduced the punishment expenditure by 1.49 MUs. This punishment pattern led to a hump-shaped relation between an individual's income and the deviation from the average cooperation of the other group members. The income was highest when the individual's investment was close to the average investment of the others. Both positive and negative deviations from the average investment decreased an individual's income.

The punishment of non-cooperators substantially increased the amount that subjects invested in the public good. In the five sessions where the punishment condition was the first treatment (Fig. 2a), the average cooperation level was much higher in the punishment condition (Wilcoxon signed rank test, five matched observations, $z = -2.023$, $P = 0.043$, two-tailed). The average investment of 94.2% of the subjects was higher in the punishment condition. In fact, the average investment in the punishment condition was higher in each session and in each period than the average investment in any of the periods and sessions of the no-punishment condition. The time trend of the average investment was also rather different in the two conditions (Fig. 2a). Although cooperation increased over time in the punishment condition, it sharply decreased in the no-punishment condition. In the final period of the punishment condition, 38.9% of the subjects contributed their whole endowment and 77.8% contributed 15 MUs or more. In the final period of the no-punishment condition, 58.9% of the subjects contributed nothing and 75.6% contributed 5 MUs or less.

A very similar pattern emerged in the sessions where the no-punishment condition came first (Fig. 2b). The average cooperation again was much higher in the punishment condition (Wilcoxon signed rank test, five matched observations, $z = 2.023$, $P = 0.043$, two-tailed). In the punishment condition, 91.4% of the subjects contributed more than in the no-punishment condition. In addition, although the average investment decreased in the no-punishment condition, it increased sharply in the punishment condition. Moreover, Fig. 2b indicates that the punishment threat was immediately effective because there was a large upwards jump in investments when the punishment opportunity was made available to the subjects. It also turns out that the sequence of the treatments had no effect on cooperation. Investments in the punishment condition were similar, irrespective of whether this condition came first or second in a session (Mann–Whitney test, $z = 0.104$, d.f. = 4, $P = 0.918$, two-tailed). The same held for the no-punishment condition (Mann–Whitney test, $z = 1.358$, d.f. = 4, $P = 0.175$, two-tailed). Thus, we can use the data of all ten sessions to compare the average investments across conditions so that the differences are significant at a much higher level (Wilcoxon signed rank test, ten matched observations, $z = 2.803$, $P = 0.005$, two-tailed).

It is not only the punishment opportunity (that is, the non-executed punishment threat) but also the actual punishment that raised cooperation levels. When a subject was punished before period 6, that subject raised investment in the next period on average by 1.62 MUs. Note, however, that this does not constitute an indirect material benefit of the act of punishment for an individual punisher, because the punishing subject never meets

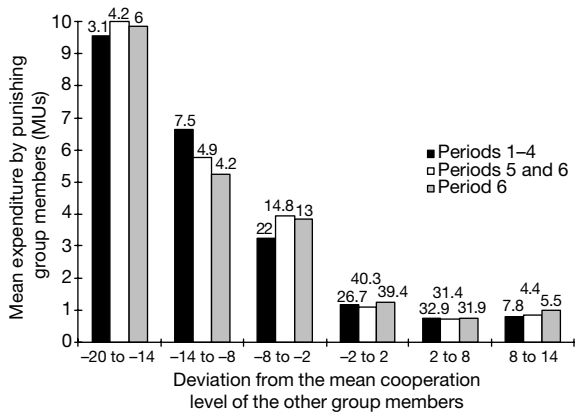

Figure 1 Mean expenditure on punishment during different time intervals as a function of the deviation of the cooperation of the punished group member from the mean cooperation of the other members. Each money unit (MU) spent on punishment reduced the income of the punished member by 3 MUs. The numbers above the bars indicate the relative frequency of the observations underlying the bars. For example, during periods 1–4, individual group members deviated between –20 and –14 MUs from the average cooperation of other group members in 3.1% of all cases.

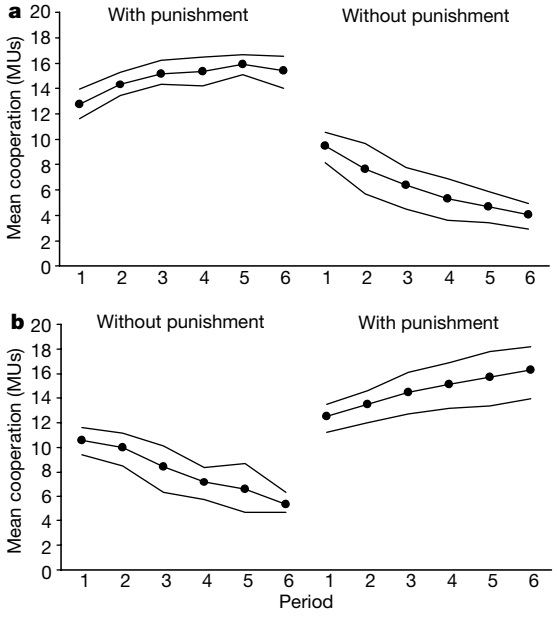

Figure 2 Time trend of mean cooperation together with the 95% confidence interval. **a**, During the first six periods, subjects have the opportunity to punish the other group members. Afterwards, the punishment opportunity is removed. **b**, During the first six periods, punishment of other group members is ruled out. Afterwards, punishment is possible.

the same subjects again. The act of punishment does provide a material benefit for the future interaction partners of the punished subject but not for the punisher. Thus, the act of punishment, although costly for the punisher, provides a benefit to other members of the population by inducing potential non-cooperators to increase their investments. For this reason, the act of punishment is an altruistic act.

Emotions as a proximate mechanism

Given the pattern of punishment, the investment behaviour of subjects seems quite rational. To avoid punishment, subjects invested in accordance with the group norm. But we wondered why subjects punish free riders in a one-shot context when this is costly. With regard to the proximate source of the punishment, negative emotions may provide an explanation. Free riding may cause strong negative emotions among the cooperators and these emotions, in turn, may trigger their willingness to punish the free riders[24,25]. If this conjecture is correct, we should observe particular emotional patterns in response to free riding. To elicit these patterns, the participants were confronted with the following two hypothetical investment scenarios after the final period of the second treatment (the numbers in brackets relate to the second scenario):

"You decide to invest 16 [5] francs to the project. The second group member invests 14 [3] and the third 18 [7] francs. Suppose the fourth member invests 2 francs to the project. You now accidentally meet this member. Please indicate your feeling towards this person."

After they had read a scenario, subjects had to indicate the intensity of their anger and annoyance towards the fourth person (the free rider) on a seven-point scale (1 = 'not at all' to 7 = 'very much'). The difference between scenarios 1 and 2 is that the other three persons in the group contribute relatively much in scenario 1 and relatively little in scenario 2. It turns out that a free rider triggered much anger among the other subjects if these subjects contributed a lot relative to the free rider (scenario 1). Forty-seven per cent of the subjects indicated an anger level of 6 or 7 and another 37% indicated an anger level of 5. If the deviation of the free rider's contribution from the other members' contribution was relatively small (scenario 2), the anger level was significantly lower (Wilcoxon signed rank test, $z = 9.636$, $P < 0.0005$) but still considerable. In this case (scenario 2), 17.4% of the subjects indicated an anger level of 6 or 7 and 80.5% indicated a level of 4 or 5 in scenario 2. This shows that the intensity of negative emotions towards a free rider varies with the deviation from the others' average contribution.

Because we were also interested in the free riders' expectation of the other members' anger, we confronted subjects with a third and a fourth hypothetical scenario (numbers in brackets relate to scenario 4):

"Imagine that the other three group members invest 14, 16 and 18 [3, 5 and 7] francs to the project. You invest 2 francs to the project and the others know this. You now accidentally meet one of the other members. Please indicate the feelings you expect from this member towards you."

In scenarios 3 and 4, the hypothetical free rider had to indicate the expected anger of the others on a seven-point scale. The anger that was expected by the free riders in scenario 3 was even greater than the actually expressed anger according to scenario 1 (Wilcoxon signed rank test, $z = 7.68$, $P < 0.0005$). In scenario 3, 74.5% of the subjects expected the anger level of others to be 6 or 7, and 22.5% expected an anger level of 5. In scenario 4, the deviation of the hypothetical free rider from the others' contribution was smaller than in scenario 3. This decrease in the deviation from the others caused significant differences in expected anger levels between scenarios 3 and 4 (Wilcoxon signed rank test, $z = 12.17$, $P < 0.0005$). Only 17.8% of the hypothetical free riders expected anger levels of 6 or 7 in scenario 4 and 80% expected levels of 4 or 5.

The low contributors in the no-punishment condition expected a higher intensity of negative emotions than the high contributors. This probably reflects the fact that the low contributors in the no-punishment condition experienced more sanctions in the punishment condition.

The same four scenarios were presented to 33 subjects that had not previously participated in our experiments, to check whether participating in the experiment affects the elicited emotions. The same emotional patterns that were expressed by our 240 experimental subjects were expressed by the 33 subjects who did not participate in our games.

Our results suggest that free riding causes strong negative emotions and that most people expect these emotions. Moreover, the above emotional pattern is consistent with the hypothesis that emotions trigger punishment for the following reasons. First, if negative emotions trigger punishment, most punishment acts would be expected to be executed by above-average contributors and imposed on below-average contributors. This is clearly the case in our experiments: 74.2% of all punishment acts follow this pattern. Second, punishment increased with the deviation of the free rider from the average investment of the other members. This is exactly what would be expected if negative emotions are the proximate cause of the punishment, because negative emotions became more intense as the free rider deviated further from the others' average investment. Third, if negative emotions cause punishment, the punishment threat is rendered immediately credible because most people are well aware that they trigger strong negative emotions when they free ride. Therefore, we should detect an immediate impact of the punishment opportunity on contributions at the switch points between the punishment and the no-punishment condition. This is what we observed. The introduction (or elimination) of the punishment opportunity led to an immediate rise (or fall) in investment (see Fig. 2). Taken together, these observations are consistent with the view that emotions are an important proximate factor behind altruistic punishment.

Our evidence has profound implications for the evolutionary study of human behaviour. In the past, human cooperation has mainly been explained in terms of kin selection, reciprocal altruism, indirect reciprocity and costly signalling. These theories focus attention on mechanisms other than altruistic punishment. By showing that altruistic punishment is a key force in the establishment of human cooperation, our study indicates that there is more at work in sustaining human cooperation than is suggested by these theories. Thus, our evidence suggests that the evolutionary study of human cooperation in large groups of unrelated individuals should include a focus on explaining altruistic punishment[3,26,27]. Moreover, because altruistic punishment occurs among genetically unrelated individuals and under conditions that rule out direct reciprocity and reputation formation, the above-mentioned theories do not readily account for altruistic punishment. □

Methods

A total of 240 undergraduate students (31% females) from the University of Zürich and the Federal Institute of Technology (ETH) voluntarily participated in the experiments. Special care was exerted to recruit students from many different disciplines to maximize the chances that the subjects had never met before. Ten experimental sessions with 24 subjects took place. Each of the 24 subjects played two 6-period public goods games: a game without a punishment opportunity and a game with a punishment opportunity. In five sessions, subjects first played the punishment treatment and then the no-punishment treatment; in the other five sessions, the treatment sequence was reversed. When subjects played the first six-period game, they did not know that another game would take place after period 6. At the beginning they were informed that the experiment would last for six periods. After period 6, subjects were told that another six-period experiment would take place and that thereafter the whole session would be over. The experiments typically lasted 60 min and on average subjects earned 39.7 Swiss francs (US$23.95, 27.2 euros) per session.

In each period of a session, the 24 subjects were randomly allocated to six groups of four subjects. The allocation of subjects to the groups ensured that, within a given treatment, no one ever met the same person more than once. In every period, the group members knew nothing about the previous cooperation and punishment decisions of the others in the

group, which ensured that subjects could not develop any kind of reputation. At the end of each period, subjects were informed about their own decisions, the decisions of the other group members, and their monetary pay-off in the current period.

At the beginning of the experiment, subjects received written instructions (available from the authors on request) that explained to them the pay-off structure of the game, the random recomposition of groups across periods, the anonymous identities of the other session members, the undisclosed history of the previous actions of their current group members, and the fact that they would be privately paid their experimental earnings at the end of the session. After subjects had read the instructions, but before the start of the experiment, we determined whether subjects understood the pay-off structure of the game: subjects had to compute their own pay-off and the pay-offs of their group members in several hypothetical examples. Every subject solved these exercises correctly. All experimental decisions were made on a computer screen using the experimental software z-tree[28]. Each of the 24 computers was located in a booth such that subjects could not see each other. To relate subjects' decisions to their dispositions, subjects filled out a personality questionnaire before and after the experiment. In addition, subjects filled out an emotions questionnaire after the experiment.

Received 5 October; accepted 5 November 2001.

1. Smuts, B. B., Cheney, D. L., Seyfarth, R. M., Wrangham, R. W., & Struhsaker, T. T. (eds) *Primate Societies* (Univ. Chicago Press, Chicago, 1987).
2. Richerson, P. & Boyd, R. in *Ideology, Warfare and Indoctrinability* (eds Eibl-Eibesfeldt, I. & Salter, F.) 71–95 (Berghan Books, New York, 1998).
3. Sober, E. & Wilson, D. S. *Unto Others: The Evolution and Psychology of Unselfish Behaviour* (Harvard Univ. Press, Cambridge, Massachusetts, 1998).
4. Boyd, R. & Richerson, P. in *Evolution and Culture* (ed. Levinson, S.) (MIT Press, Cambridge, Massachusetts, in the press).
5. Hamilton, W. D. Genetical evolution of social behavior I and II. *J. Theor. Biol.* **7,** 1–52 (1964).
6. Trivers, R. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46,** 35–57 (1971).
7. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211,** 1390–1396 (1981).
8. Axelrod, R. *The Evolution of Cooperation* (Basic Books, New York, 1984).
9. Nowak, M. A., May, R. M. & Sigmund, K. The arithmetics of mutual help. *Sci. Am.* **272,** 76–81 (1995).
10. Alexander, R. D. *The Biology of Moral Systems* (Aldine de Gruyter, New York, 1987).
11. Nowak, M. A. & Sigmund, K. The dynamics of indirect reciprocity. *J. Theor. Biol.* **194,** 561–574 (1998).
12. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393,** 573–577 (1998).
13. Lotem, A., Fishman, M. A. & Stone, L. Evolution of cooperation between individuals. *Nature* **400,** 226–227 (1999).
14. Wedekind, C. & Milinski, M. Cooperation through image scoring in humans. *Science* **288,** 850–852 (2000).
15. Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. Lond. B* **268,** 745–753 (2001).
16. Zahavi, A. The cost of honesty (further remarks on the handicap principle). *J. Theor. Biol.* **67,** 603–605 (1977).
17. Zahavi, A. Altruism as a handicap—the limitations of kin selection and reciprocity. *J. Avian. Biol.* **26,** 1–3 (1995).
18. Gintis, H., Smith, E. & Bowles, S. Costly signalling and cooperation. *J. Theor. Biol.* **213,** 103–119 (2001).
19. Clutton-Brock, T. H. & Parker, G. A. Punishment in animal societies. *Nature* **373,** 209–216 (1995).
20. Axelrod, R. An evolutionary approach to norms. *Am. Pol. Sc. Rev.* **80,** 1095–1111 (1986).
21. Heckathorn, D. D. Collective action and the second-order free-rider problem. *Ration. Soc.* **1,** 78–100 (1989).
22. Henrich, J. & Boyd, R. Why people punish defectors. *J. Theor. Biol.* **208,** 79–89 (2001).
23. Ostrom, E., Walker, J. & Gardner, R. Covenants with and without a sword: self governance is possible. *Am. Pol. Sci. Rev.* **86,** 404–417 (1992).
24. Hirshleifer, J. in *The Latest on the Best: Essays on Evolution and Optimality.* (ed. Dupré, J.) (MIT Press, Cambridge, Massachusetts, 1987).
25. Frank, R. *Passions within Reason: The Strategic Role of the Emotions* (Norton, New York, 1988).
26. Gintis, H. Strong reciprocity and human sociality. *J. Theor. Biol.* **206,** 169–179 (2000).
27. Sigmund, K., Hauert, C. & Nowak, M. A. Reward and punishment. *Proc. Natl Acad. Sci. USA* **98,** 10757–10761 (2001).
28. Fischbacher, U. *Z-tree: Zürich Toolbox for Readymade Economic Experiments* Working Paper No. 21 (Institute for Empirical Research in Economics, Univ. Zürich, 1999).

Acknowledgements

Support by the MacArthur Foundation Network on Economic Environments and the Evolution of Individual Preferences and Social Norms, and the EU-TMR Research Network ENDEAR is gratefully acknowledged. We also thank R. Boyd, A. Falk, U. Fischbacher, H. Gintis and J. Henrich for comments, and M. Näf and D. Reding for research assistance. We are particularly grateful to U. Fischbacher for writing the computer software.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.F. (e-mail: efehr@iew.unizh.ch).