Constraining free-riding in public goods games: Designated solitary punishers can sustain human cooperation.

Running head: Solitary punishers constrain free-riding

Rick O'Gorman[1], Joseph Henrich[2], & Mark van Vugt[3]

1. Psychology Group, Sheffield Hallam University, Collegiate Crescent Campus, Sheffield, S10 2BP, UK

2. Psychology and Economics Departments, University of British Columbia, Vancouver, V6T 1Z4, Canada

3. Psychology Department, University of Kent, Canterbury, CT2 7NP, UK

Corresponding author: Rick O'Gorman, Psychology Group, Sheffield Hallam University, Collegiate Crescent Campus, Sheffield, S10 2BP, UK, Tel: +44 114 225 5788, Fax: +44 114 225 4449, rogorman@alumni.binghamton.edu

Number of words: 5693 (excludes title page)

Number of figures: 4

Number of tables: 3

Summary

Much of human cooperation remains an evolutionary riddle. Unlike other animals, people frequently cooperate with non-relatives in large groups. Evolutionary models of large-scale cooperation require not just incentives for cooperation, but also a credible disincentive for free-riding. Various theoretical solutions have been proposed and experimentally explored, including reputation monitoring and diffuse punishment. Here, we empirically examine an alternative theoretical proposal: Responsibility for punishment can be borne by one specific individual. This experiment shows that allowing a single individual to punish increases cooperation to the same level as allowing each group member to punish and results in greater group profits. These results suggest a potential key function of leadership in human groups and provide further evidence supporting that humans will readily and knowingly behave altruistically.

Key terms: Cooperation, free-riding, punishment, altruism, leadership

1. Introduction

In recent years, there has been a spate of papers providing evidence for various mechanisms to coax cooperation out of groups of individuals (Rockenbach & Milinski, 2006; Sigmund 2007). It is to state the obvious that humans can cooperate readily in extraordinary numbers (Smirnov et al 2007) and that this cooperation often provides public goods, despite the risk of free-riding (Andreoni 1988; Fehr & Fischbacher 2003). Much of the recent empirical work on the puzzling aspects of human cooperation have focused on testing evolutionary models of diffuse or altruistic punishment (Boyd & Richerson 1992; Boyd et al 2003; Henrich & Boyd 2001), in which many individuals share the burden of punishing non-cooperators (Fehr & Fischbacher 2003; Fehr & Gächter 2002; Fehr et al 2002; Sober & Wilson 1998).

However, since recent work has shown a lack of motivation for costly punishment in some otherwise cooperative societies (Henrich et al 2006)--perhaps because the solutions have not addressed the problem of second-order free-riding--and a possible taste for countervailing anti-social punishment (Hermann et. al. 2008), it seem plausible that different mechanisms may stabilize cooperation in different ways in different populations. We explore a solution to *n*-person cooperation in which a designated individual is responsible for punishment. Over the course of human evolution individuals in groups capable of motivating cooperation would have gained an adaptive advantage. Observed hunter-gatherer groups adopt various mechanisms to ensure cooperation, and leadership is one such mechanism that both integrates with humanity's primate heritage and offers a mechanism for groups to coordinate activity (Boehm 1999; Brown 1991; Van Vugt, 2006). Models in economics (Hirshleifer & Rasmusen 1989) and evolutionary biology (Boyd &

Richerson 1992) indicate that evolution can favour a single punisher per social group and that the actions of this one punisher can efficiently galvanize group cooperation. This solution is particularly interesting since it lacks the second-order free-rider problem--which has been the central focus of much theoretical effort--and it avoids the problem of uncoordinated over-punishment.

Our experimental findings confirm that (1) when placed in the sole punisher role individuals will punish sufficiently to sustain cooperation, (2) others will respond by increasing cooperative contributions, and (3) a single punisher can sustain levels of cooperation comparable to that maintained by diffuse punishment (Fehr & Gächter 2002; Ostrom et al 1994; Yamagishi 1986) and at more profitable levels, since punishing efforts are not unnecessarily duplicated. Such findings suggest that in the smaller-scale societies that have dominated human evolutionary history (as well as in the smaller groups of contemporary societies) the single punisher solution may have been an important means of maintaining cooperation. In such groups, single punishers may even be a superior mechanism, compared to diffuse punishment systems.

2. Methods

(a) *Participants*

136 participants (35% male) who were undergraduate students from the University of Kent at Canterbury were recruited from across the campus by way of a job advertisement service. Six experimental sessions took place with 20-24 participants per session. The sessions lasted approximately one hour and the average earning for participants was UK£5.47. Each MU earned during the session equated to UK£.01.

(b) *Design and procedure*

Initially, participants were informed, by way of a projected presentation, of the procedure of the experiment (including that assignment to groups was random and occurred each round, interactions were anonymous, the amount of endowment, how it could be invested, how payoffs were allocated, how they would be paid), examples of different contribution patterns and the corresponding payoffs, and the use of the computer software. Participants were not informed at the beginning of the first segment that there were two segments. After the presentation of the instructions, participants were tested on their understanding of the payoff procedure. All participants showed satisfactory comprehension.

For those in punishment conditions, further instruction was provided prior to the commencement of the second segment while those continuing with a second control condition received a brief refresher. Instructions relating to the making of deductions did not make any suggestion as to how such deductions could be used, or whether they should be used. Participants were simply informed that such deductions would be possible for the second segment and it was explained how to make such deductions, should participants wish to use such a facility. If a participant queried the purpose, then he or she was simply told that it was an option that would be available and it was up to him or here how it could be used.

We used a modified methodology (Fehr & Gächter 2002) of a public goods experiment that had real monetary earnings at stake run on networked PCs using z-Tree software (Fischbacher 2007). Participants all completed a two-segment experiment with an initial no-punishing control segment followed by a second segment of either a further control condition (no-punishment), a condition with punishment permitted for all group

members (all-punishment), or a condition with only one individual permitted to punish (one-punishment); therefore all participants acted as their own controls and partook in only one of three conditions. We did not counterbalance as, firstly, Fehr and Gächter (2002) showed that there was no order effect for not-punishing versus punishing and, secondly, our focus was on comparison between the two punishment conditions.

In all conditions, participants played the same public goods game: Assigned to groups of four, participants were allocated an endowment of 20 monetary units (MUs), of which they could invest any amount into a group fund and retain the remainder. Each MU invested in the group fund yielded a payoff of 0.5MU to each group member, irrespective of who invested. That is, each MU invested in the group fund was doubled and then divided four ways. Thus, participants would always be better off contributing nothing to the group fund as the return was less than the investment. However, if every member invested their full endowment, then each member would earn 40 MUs, a profit of 20.

Each round, groups were randomly formed so that participants never knew with whom they were interacting ("stranger protocol" in the economics literature), thus controlling for reputation and reciprocity effects. All interactions were anonymous. Investment decisions were made simultaneously, after which information was provided on the investments of other group members. In the second-segment punishment conditions, individuals could simultaneously make deductions from each other by paying a fee, drawn from their earnings for that round, up to a fee maximum of 10 MUs per punished member (the deduction was equivalent to three times the fee). For the one-punishment condition, one member per group was randomly selected after each investment phase to make deductions, whereas in the all-punishment condition, all individuals could make deductions.

We conducted the public goods game for six rounds in each condition, so that participants played a total of twelve rounds over two segments to avoid one-shot effects and to examine participants' behaviour over a series of games. With each participant acting as his or her own control and with a fixed order, we could compare between conditions.

During the experiment, participants received no information other than of the contributions made by each of the other group members to the group fund and, in the punishing conditions after punishing occurred, of the level of deductions made from their own account only. Participants were located in a large computer laboratory and were spaced apart such that no-one could see another participant's screen. After completing the public goods games, participants completed online and paper questionnaires to assess their attitudinal and emotional responses to the experiment and their interactions in the games, group identity, and a number of other measures not reported here.

3. Results

The average contribution made by participants across all sessions and rounds was 8.28 MUs (SD=6.55). For analysis, we used Generalized Estimating Equations, available in SPSS 15, which utilizes robust (Huber-White) errors to correct for lack of independence in the data. Because participants interacted with each other within sessions, this represents a conservative approach to analysis (we also performed a non-parametric analysis, which yielded qualitatively similar results; however GEE allows for more powerful analysis and is what we report here). We present our analysis firstly of the contribution data (segment one followed by segment two), then of the profit data and finally of the punishing data.

3.1 Analysis of contributions

GEEs for the contribution data used a first-order autoregressive working correlation matrix, due to correlations between adjacent rounds' contributions and a normal/identity link. The data from segment two of the study concerns our primary hypothesis that the one-punishment condition would increase contributions above the control condition (see Fig.1 for mean contributions and 95% confidence intervals for the three conditions over the six rounds). We examined the effects of condition, round and sex on contributions. There is a main effect on segment two contributions (Wald $\chi^2 = 10.41$, d.f. = 2, $p = .005$). Regression values (derived from the GEE model, see Table 1) show that both all-punishment and one-punishment differ significantly from the control group (all-punishment v control: B = 6.20, S.E. = 1.32, $p < .001$; one-punishment v control: B = 5.47, S.E. = 1.19, $p < .001$) while all-punishment and one-punishment do not appear to significantly differ (B = -.73, S.E. = 1.32, $p = .579$; obtained by switching the reference category from control to all-punishment).

Additionally, a main effect for round approaches significance (Wald $\chi^2 = 10.74$, d.f. = 5, $p = .057$) and there is an interaction between manipulation and rounds (Wald $\chi^2 = 25.29$, d.f. = 10, $p = .005$), reflecting the decrease in contributions in the no-punishment condition in contrast to the more stable contributions in the other two conditions (see Fig.1). Contributions in the control condition decreased significantly across the six rounds in segment two (rounds regressed on contributions with robust errors, B = -1.08, S.E. = .22, $p <.001$) whereas contributions in the two punishing conditions remained relatively constant (all-punishment B = -.12, S.E. = .21, $p = .570$; one-punishment B = .03; S.E. = .19, $p = .868$). There is no effect for sex, nor is there an interaction ($p > .370$). Our findings support the hypothesis that, under these conditions, a single individual operating as the sole

punisher in a group can improve contributions relative to a control condition without punishment and matches the effect produced by allowing everyone to punish.

We should note that there are differences between conditions in contributions (Wald χ^2 = 19.59, d.f. = 2, $p < .001$), possibly due to participants attending with an understanding of the experiment, but these initial differences disappear after the six rounds (round one: Wald χ^2 = 16.72, d.f. = 2, $p < .001$; round six: Wald χ^2 = 1.80, d.f. = 2, $p = .408$). However, this change is not reflected in a significant interaction (Wald χ^2 = 4.68, d.f. = 10, $p = .912$), though there is a main effect for round (Wald χ^2 = 33.09, d.f. = 5, $p < .001$). Finally, there is no difference due to sex of participant (Wald χ^2 = 2.91, d.f. = 1, $p = .088$), nor did sex interact with either condition or round ($p > .711$). The lack of a significant difference between conditions in round six of segment one suggests that any initial differences in contribution levels between conditions had been eliminated by the end of segment one, but to control for differences in baseline contribution dispositions, we used participants' average contributions in segment one as a covariate in the analysis of segment two data.

3.2 Analysis of profits

Differences in segment one profits due to condition and round necessarily follow contribution differences in the same study segment and so, not surprisingly, are significant (condition: Wald χ^2 = 25.68, d.f. = 2, $p < .001$; round: Wald χ^2 = 33.68, d.f. = 5, $p < .001$), though there is no effect for sex or interactions. As above, for analysis of the profit data from segment two (see Fig.2 for mean profits and 95% confidence intervals), we use segment one contributions as a covariate. There is a main effect for condition (Wald χ^2 = 144.79, d.f. = 2, $p < .001$) and an interaction between condition and round (Wald χ^2 = 42.10, d.f. = 10, $p < .001$), though no main effects for round or sex, nor are there interaction

effects. Regression values (as earlier, derived from the GEE model, see Table 2) show that all three conditions differ (all-punishment v control: B = -11.55, S.E. = 2.03, $p < .001$; one-punishment v control: B = -5.48, S.E. = 1.79, p = .002; all-punishment v one-punishment: B = 6.07, S.E. = 2.32, p = .009; the latter is again obtained by switching the reference category from control to all-punishment).

The lower mean values for the punishment conditions is primarily due to the cost of punishing and to deductions, relative to the control condition. However, it is worth noting that, whereas the slopes of the punishment conditions appear stable relative to rounds (rounds regressed on contributions with robust errors, all-punishment B = .01, S.E. = .46, p = .980; one-punishment B = .21; S.E. = .34, p = .542), the control condition's slope is not (B = -1.08, S.E. = .23, $p <.001$), suggesting that both punishment conditions would be likely to be more profitable than the control condition in the long run. Importantly, the one-punishment condition has an advantage over the all-punishment condition due to lower total costs incurred by group members and this is reflected in its higher profit levels (see Fig.2).

3.3 Analysis of punishment

Looking at punishing, overall participants in the all-punishment condition punished on 38.9% of opportunities to do so whereas punishers in the one-punishment condition did so on 56.6% of opportunities. Per round, the proportion of participants who punished in the one-punishment condition (i.e., punished at least once) was greater than the proportion in the all-punishment condition (see Fig.3a). Fewer participants were punished in the one-punishment condition (Fig.3b) but that condition's punishers made greater deductions (Fig.3c), although the total incurred punishments were not consistently harsher in either condition (Fig.3d).

When we examined possible factors that affected punishment behaviour, we found that punishers in the all-punishment condition and the one-punishment condition appear to be influenced by different factors. Using a GEE approach (using a gamma/log link, due to the positively skewed data), we examined separately for the two punishment conditions the relationship of punishment levels with the contribution of potential punishers and targets, deviations from group and session means by targets, and rounds. Results for the three different measures of targets' contributions are similar, due to these measures being highly correlated (r's > .90). Thus, we focus here on our analysis for actual contributions, as that analysis produced the (marginally) stronger results. Examining the all-punishment condition first (see Table 3 for GEE regression parameter estimates), we found that higher punishments resulted from lower contributions by the target (B = -.03, S.E. = .01, $p < .001$), as might be expected, but, when we added the interaction between the sender's and target's contributions, then the target's contributions are no longer significant but both the sender's contribution (B = .03, S.E. = .01, p = .008) and the interaction are significant (B = -.005, S.E. = .001, $p < .001$). In contrast, adding the interaction term in the one-punishment condition results in no significant predictors. Without the interaction term, the target's contribution is a significant predictor (B = -.02, S.E. = .01, $p < .033$).

Thus, it appears that while participants in diffuse punishment situations attend to both their own contribution and that of the target, perhaps using their own contributions to guide their decision on whether to punish, those in the solitary punisher condition attend only to the contributions of the target, possibly focussed solely on whether contributions are maximally beneficial for the group, in which case any deviation from a full contribution represents an undesirable shortfall. Fig.4 shows that for both punishment conditions, lower target

contributions are associated with higher punishment, though this pattern is clearer for those in the all-punishment condition (panel a).

4. Discussion

Individual contributions were significantly higher when punishment was available as an option, with participants responding as effectively to a single individual as to all group members making deductions. Our results suggest that a single-punisher successfully enhances and stabilizes group contributions, while doing so more profitably than in the all-punishment condition. As punishment costs are lost to the system, punishments by a single punisher are more coordinated and thus reduce inefficient losses. It is important to note that the success of punishing in this study (in either punishment condition) is facilitated by the 1:3 ratio of the cost of punishing for the punisher to the cost for the target. While this is a common ratio in this methodology, studies have shown that lower ratios tend to not produce punishing behaviour sufficient to sustain cooperation (Burnham & Johnson 2005; Nikiforakis & Normann 2008; Yamagishi 1986). However, we do not view this as an unnecessary stumbling block. Asymmetrical impacts of punishment can be readily achieved in the real world, for example, through the use of a weapon or social support.

The pattern of punishment for contribution levels suggests that lower contributions tend to incur greater levels of punishment. As Carpenter and Matthews (2008) argue, it appears that punishers are more focussed on actual contribution levels rather than deviations from the group (or session) mean, per se. However, actual strategies in anonymous games inevitably are likely to be complex, reflecting the fact that individuals vary in their cooperative intent (Van Lange 1999) and thus how they respond to both being able to ‘punish’ and being ‘punished’. Further in-depth examination of participants

strategies, motives and goals is needed. Somewhat unexpectedly, more participants in the one-punishment condition punished more often and more harshly than in the all-punishment condition, incurring greater personal costs. This finding supports the claim for an evolved altruistic proclivity in humans to punish free-riders to the benefit of the group in the absence of reputation enhancement (Fehr & Gächter 2002). That individuals punished knowing that they were anonymous and, when in the one-punishment condition, the sole potential punisher, questions the notion that humans do not altruistically punish. Humans should be evolved to determine how costly actions might impact their reputation. It is also possible that diffusion of responsibility was reduced by having a sole punisher (Latané & Darley 1970), though anonymity and costs mean that this cannot solely explain the behaviour.

These findings may have an important implication for the study of cooperation and the functions of leadership in humans. As noted earlier, large scale cooperation in human groups (beyond the hunter-gather level) represents an evolutionary puzzle. Diffuse punishment does not fully solve this issue because of the iterated problem of second-order free-riders. A system with a single designated punisher can potentially avoid this problem because there is clearer accountability. In human groups, leaders often fulfil the role of designated punishers (Diamond 1997; Heizer 1978; Krackle 1978). Moreover, some form of leadership, even if only ephemeral (Johnson & Earle 2000; Steward 1938), is a human universal and readily emerges in ad-hoc laboratory groups (Van Vugt 2006).

Of course, such a leadership role is potentially costly to the individual who occupies it. There is both the energy budget of punishing, and the incumbent costs of self-defence by the target or retaliation. Why would individuals take on this role? There may be

compensatory benefits for acting as a leader. Some individuals more readily fulfil this role than others based on heritable differences in personality (Hogan et al 1994). In human societies, leaders acquire status and prestige (Van Vugt 2006), which may translate into increased reproductive success (Fieder et al 2005; Henrich & Gil-White 2001). Alternatively, group-level selection could facilitate leadership emergence, either by genetic or cultural mechanisms (Richerson & Boyd 2004; Sober & Wilson 1998). Groups often favour altruists for the leader role (Hardy & Van Vugt 2006; Milinski et al 2002). Competition between rival groups results in selective pressure for its adoption culturally or its evolution, genetically. If all participants can punish each other, such situations risk deteriorating into retaliatory actions that do not just reduce benefits from joint activities but damage the integrity of the group (Denant-Boemont et al 2007; Nikiforakis 2008). A designated punisher avoids these risks.

The issue of anonymity and the consequential inability of punished individuals to retaliate represent a constraint on our argument that we provide evidence for leadership to function as a constraint on free-riding. If retaliation were possible, a single punishing individual would be less costly to retaliate against than a set of punishers. However, in this study, we seek only to demonstrate that leadership could fulfil such a function successfully. In reality, a leader is not just one individual but represents the pinnacle of a social structure. Thus, although responsibility may lie with one individual to act, such actions nonetheless, by virtue of the role, carry the support of the group, or at least a majority. Additionally, the actual form of punishment varies substantially, and indeed a leader may not need to be the individual to actually impose the punishment, as immortalized by Tony Soprano and as is very familiar to anyone working in an institution that has punishment capabilities.

In the present study, the random selection of punishers in the one-punishment condition served as a means to impose the role on individuals to control for other confounds. Nonetheless, future studies would do well to attend to more realistic exploration of the role of leaders as punishers. One interesting follow-up would be to examine a series of experimental rounds, allowing participants either to experience different regimes (no punishing, diffuse punishing, single punisher) or gain information on the performance of different regimes, and choose which system to play under. This could further demonstrate the willingness (or not) of individuals to operate under a designated punisher (leader) system. Related, research documenting cross-cultural variation in costly punishing (Henrich et al 2006) suggests our findings may be constrained and it would be worthwhile to consider whether punishing through leadership is a cultural universal. The potential impact of retailiation also warrants consideration.

In smaller-scale human societies prestigious leaders can galvanize the trace of larger-scale cooperation (Johnson 2003). At least in some circumstances, individuals respond as effectively to a single punishing individual as they do to a more general punitive environment without obvious negative reactions. Consistent with existing theoretical work (Boyd & Richerson 1992), our research suggests that human psychology may have evolved to recognize situations in which a single motivated leader can enforce cooperation (Van Vugt, Hogan, & Kaiser 2008).

References

Andreoni, J. 1988. Why free ride? Strategies and learning in public good experiments. *J. Public Econ.* **37**, 291-304.

Boehm, C. 1999 *Hierarchy in the Forest*. Cambridge, MA: Harvard University Press.

Boyd, R. & Richerson, P.J. 1992 Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171-195.

Boyd, R., Gintis, H., Bowles, S. & Richerson, P.J. 2003 The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. USA* **100**, 3531-3535.

Brown, D. 1991 *Human Universals*. Boston, MA: McGraw-Hill.

Burnham, T.C. & Johnson, D.D.P. 2005 The biological and evolutionary logic of human cooperation. *Analyse & Kritik* **27**, 113-135.

Carpenter, J.P. & Matthews, P.H. 2008 What norms trigger punishment. Working paper obtained from http://community.middlebury.edu/~jcarpent/papers.html on June 20th, 2008.

Denant-Boemont, L., Masclet, D. & Noussair, C.N. 2007 Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Econ. Theor.* **33**, 145-167.

Diamond, J. 1997 *Guns, Germs, and Steel: The Fates of Human Societies*. New York: Norton.

Fehr, E. & Fischbacher, U. 2003 The nature of human altruism. *Nature* **425**, 785-791.

Fehr, E. & Gächter, S. 2002 Altruistic punishment in humans. *Nature* **415**, 137-140.

Fehr, E., Fischbacher, U. & Gächter, S. 2002 Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum. Nature* **13**, 1-25.

Fieder, M. et al 2005 Status and Reproduction in Humans: New Evidence for the Validity of Evolutionary Explanations on Basis of a University Sample. *Ethology* **111**, 940-950.

Fischbacher, U. 2007 z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Economics* **10,** 171-178.

Hardy, C. & van Vugt, M. 2006 Nice Guys Finish First: The Competitive Altruism Hypothesis. *Pers. Soc. Psychol. Bull.* **32,** 1402-1413.

Heizer, R. 1978 (ed.) *Handbook of North American Indians: California, 8.* Washington, DC: Smithsonian Institution.

Henrich, J. & Boyd, R. 2001 Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *J. Theor. Biol.* **208**, 79-89.

Henrich, J. & Gil-White, F.J. 2001 The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evol. Hum. Behav.* **22**, 165-196.

Henrich, J. et al 2006 Costly punishment across human societies. *Science* **312**, 1767-1770.

Hirshleifer, D. & Rasmusen, E. 1989 Cooperation in a repeated prisoners' dilemma with ostracism. *J. Econ. Behav. Organ.* **12**, 87-106.

Hogan, R., Curphy, G.J. & Hogan, J. 1994 What we know about leadership. *Am. Psychol.* **49,** 493-504.

Johnson, A. 2003 *Families of the Forest: Matsigenka Indians of the Peruvian Amazon.* Berkeley: University of California.

Johnson, A. & Earle, T. 2000 *The Evolution of Human Societies*. Stanford: Stanford University Press.

Krackle, W.H. 1978 *Force and Persuasion: Leadership in an Amazonian Society.* Chicago: University of Chicago Press.

Latané, B. & Darley, J. M. 1970 *The unresponsive bystander: Why doesn't he help?* New York: Appleton-Crofts.

Milinski, M., Semmann, D. & Krambeck, H. 2002 *Proc. R. Soc. London B* **269**, 881-883.

Nikiforakis, N. 2008 Punishment and counter-punishment in public good games: Can we really govern ourselves? *J Public Econ* **92**, 91–112.

Nikiforakis, N. & Normann, H.T. 2008 A Comparative Statics Analysis of Punishment in Public Goods Experiments. *Exp. Economics*, in press.

Ostrom, E.R., Gardner, R. & Walker, J.M. 1994 *Rules, Games, and Common-pool Resources.* Ann Arbor, MI: University of Michigan Press.

Richerson, P.J. & Boyd, R. 2004 *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago: University of Chicago Press.

Rockenbach, B. & Milinski, M. 2006 The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444**, 718-723. doi:10.1038/nature05229.

Sigmund, K. 2007 Punish or perish? Retaliation and collaboration among humans. *Trends Ecol Evol* **22**, 593-600.

Smirnov, O., Arrow, H., Kennett, D. & Orbell, J. 2007 Ancestral War and the Evolutionary Origins of "Heroism". *J. Politics* **69**, 927-940.

Sober, E. & Wilson, D.S. 1998 *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Steward, J. 1938 *Basin-Plateau Aboriginal Sociopolitical Groups*. Washington DC: Bureau of American Ethnology.

Van Lange, P. A. M. 1999 The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. J *Pers. Soc. Psychol.* **77**, 337-349.

Van Vugt, M. 2006 Evolutionary Origins of Leadership and Followership. *Pers. Soc. Psychol. Rev.* **10**, 354-371.

Van Vugt, M. & Van Lange, P.A.M. 2006 Psychological adaptations for prosocial behavior: The altruism puzzle. In *Evolution and Social Psychology* (eds. M. Schaller, D. Kenrick & J. Simpson), pp. 237-261. New York: Psychology Press.

Van Vugt, M., Hogan, R., & Kaiser, R. (2008) Leadership, followership, and evolution: Some lessons from the past. *American Psychologist* **63**, 182-196.

Van Vugt, M., Jepson, S.F., Hart, C.M. & De Cremer, D. 2004 Autocratic leadership in social dilemmas: A threat to group stability. *J. Exp. Soc. Psychol.* **40**, 1-13.

Yamagishi, T. 1986 The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51,** 110-116.

Acknowledgements We acknowledge support by the British Academy for this work. We would like to thank anonymous reviewers for helpful comments.

Author contributions All authors contributed equally to this work. The authors declare that they have no competing financial interests.

Table 1. GEE parameter estimates for regression of contribution levels on punishment condition, sex and round in study segment 2. Step 1 consists of entering the main factors, step 2 introduces interaction terms. Reference categories were the control condition, round 6 and female.

Parameter	B	Std. Error	Wald Chi-Square	Sig.
Step 1				
Intercept	.689	1.0917	.398	.528
All-punishment	2.846	.8454	11.338	.001
One-punishment	2.107	.7435	8.031	.005
Round 1	1.824	.6897	6.991	.008
Round 2	1.875	.6013	9.724	.002
Round 3	1.434	.5867	5.972	.015
Round 4	.824	.5541	2.209	.137
Round 5	.324	.5763	.315	.575
Sex	.624	.7447	.701	.402
Seg1 mean contrib.	.527	.1008	27.356	.000
Step 2				
Intercept	-1.814	1.2846	1.994	.158
All-pun	6.204	1.3195	22.106	.000
One-pun	5.469	1.1918	21.060	.000
Round 1	6.271	1.3581	21.324	.000
Round 2	5.208	1.3078	15.856	.000
Round 3	4.551	1.2976	12.299	.000
Round 4	2.176	1.0505	4.291	.038
Round 5	2.283	1.2562	3.302	.069
Sex	1.905	1.4251	1.786	.181
Seg1 mean contrib.	.526	.1004	27.442	.000
All-pun * Round 1	-5.971	1.7085	12.216	.000
All-pun * Round 2	-4.502	1.4073	10.233	.001
All-pun * Round 3	-3.653	1.3486	7.338	.007
All-pun * Round 4	-2.555	1.2618	4.100	.043
All-pun * Round 5	-3.261	1.3781	5.599	.018
One-pun * Round 1	-6.027	1.5423	15.272	.000
One-pun * Round 2	-4.927	1.5063	10.699	.001
One-pun * Round 3	-3.678	1.5370	5.726	.017
One-pun * Round 4	-1.604	1.4033	1.306	.253
One-pun * Round 5	-1.796	1.4854	1.462	.227
All-pun * Sex	-.176	1.7709	.010	.921
One-pun * Sex	-1.281	1.7622	.529	.467
Round 1 * Sex	-1.638	1.4548	1.267	.260

Round 2 * Sex	-.817	1.1526	.502	.478
Round 3 * Sex	-2.156	1.1448	3.547	.060
Round 4 * Sex	-.020	1.1253	.000	.985
Round 5 * Sex	-.935	1.1631	.646	.421

Table 2. GEE parameter estimates for regression of profit levels on punishment condition, sex and round in study segment 2. Step 1 consists of entering the main factors, step 2 introduces interaction terms. Reference categories were the control condition, round 6 and female.

Parameter	B	Std. Error	Wald Chi-Square	Sig.
Step 1				
Intercept	30.003	1.4460	430.500	.000
All-punishment	-14.409	1.2864	125.468	.000
One-punishment	-7.608	.8449	81.078	.000
Round 1	1.371	1.0850	1.597	.206
Round 2	2.507	1.1470	4.779	.029
Round 3	1.287	1.1432	1.267	.260
Round 4	1.129	.9420	1.436	.231
Round 5	1.217	1.0409	1.367	.242
Sex	-1.429	.9536	2.247	.134
Seg1 mean contrib.	-.318	.1086	8.589	.003
Step 2				
Intercept	28.498	1.5377	343.479	.000
All-punishment	-11.553	2.0340	32.263	.000
One-punishment	-5.482	1.7864	9.417	.002
Round 1	5.023	1.4999	11.215	.001
Round 2	5.166	1.6530	9.769	.002
Round 3	3.039	1.5904	3.652	.056
Round 4	2.700	1.2898	4.381	.036
Round 5	1.420	1.4711	.932	.334
Sex	-1.945	1.7971	1.171	.279
Seg1 mean contrib.	-.320	.1083	8.708	.003
All-pun * Round 1	-9.657	2.5962	13.835	.000
All-pun * Round 2	-.811	2.6926	.091	.763
All-pun * Round 3	-1.466	2.7361	.287	.592
All-pun * Round 4	-1.243	2.0978	.351	.554
All-pun * Round 5	-4.287	2.5091	2.919	.088
One-pun * Round 1	-2.791	2.2027	1.605	.205
One-pun * Round 2	-6.631	2.5952	6.530	.011
One-pun * Round 3	-5.171	2.4557	4.433	.035
One-pun * Round 4	-2.371	2.2098	1.151	.283
One-pun * Round 5	2.648	2.2341	1.404	.236
All-pun * Sex	.359	2.3883	.023	.881
One-pun * Sex	.639	1.8383	.121	.728
Round 1 * Sex	1.086	2.2925	.224	.636

Round 2 * Sex	-.727	2.5144	.084	.773
Round 3 * Sex	1.141	2.3496	.236	.627
Round 4 * Sex	-1.163	1.9170	.368	.544
Round 5 * Sex	.947	2.2605	.176	.675

Table 3. GEE parameter estimates for regression of imposed punishment on sender's and target's contributions and round in study segment 2. Round 6 was the reference category for the round factor.

	All-punishment				One-punishment			
Parameter	B	Std. Error	Wald Chi-Square	Sig.	B	Std. Error	Wald Chi-Square	Sig.
Step 1								
Intercept	1.102	.1457	57.243	.000	.696	.2089	11.113	.001
Sender's contrib.	-.002	.0111	.023	.880	-.003	.0086	.104	.747
Target's contrib.	-.033	.0093	12.372	.000	-.017	.0079	4.556	.033
Round 1	.126	.1185	1.129	.288	-.139	.1342	1.072	.301
Round 2	-.144	.1444	.989	.320	.111	.2443	.208	.649
Round 3	-.072	.1485	.233	.629	.143	.2434	.345	.557
Round 4	-.109	.0974	1.245	.265	.049	.2398	.041	.839
Round 5	.049	.1445	.113	.737	-.166	.2139	.601	.438
Step 2								
Intercept	.799	.1659	23.205	.000	.680	.2232	9.279	.002
Sender's contrib.	.033	.0124	6.971	.008	-.001	.0130	.005	.942
Target's contrib.	.002	.0112	.032	.859	-.015	.0126	1.421	.233
Round 1	.131	.1130	1.336	.248	-.139	.1338	1.085	.298
Round 2	-.121	.1320	.841	.359	.113	.2436	.214	.644
Round 3	-.071	.1267	.316	.574	.143	.2430	.347	.556
Round 4	-.113	.0913	1.530	.216	.049	.2397	.042	.837
Round 5	.079	.1406	.312	.576	-.167	.2136	.610	.435
Sender's contrib. * Target's contrib.	-.005	.0011	15.829	.000	.000	.0010	.052	.820

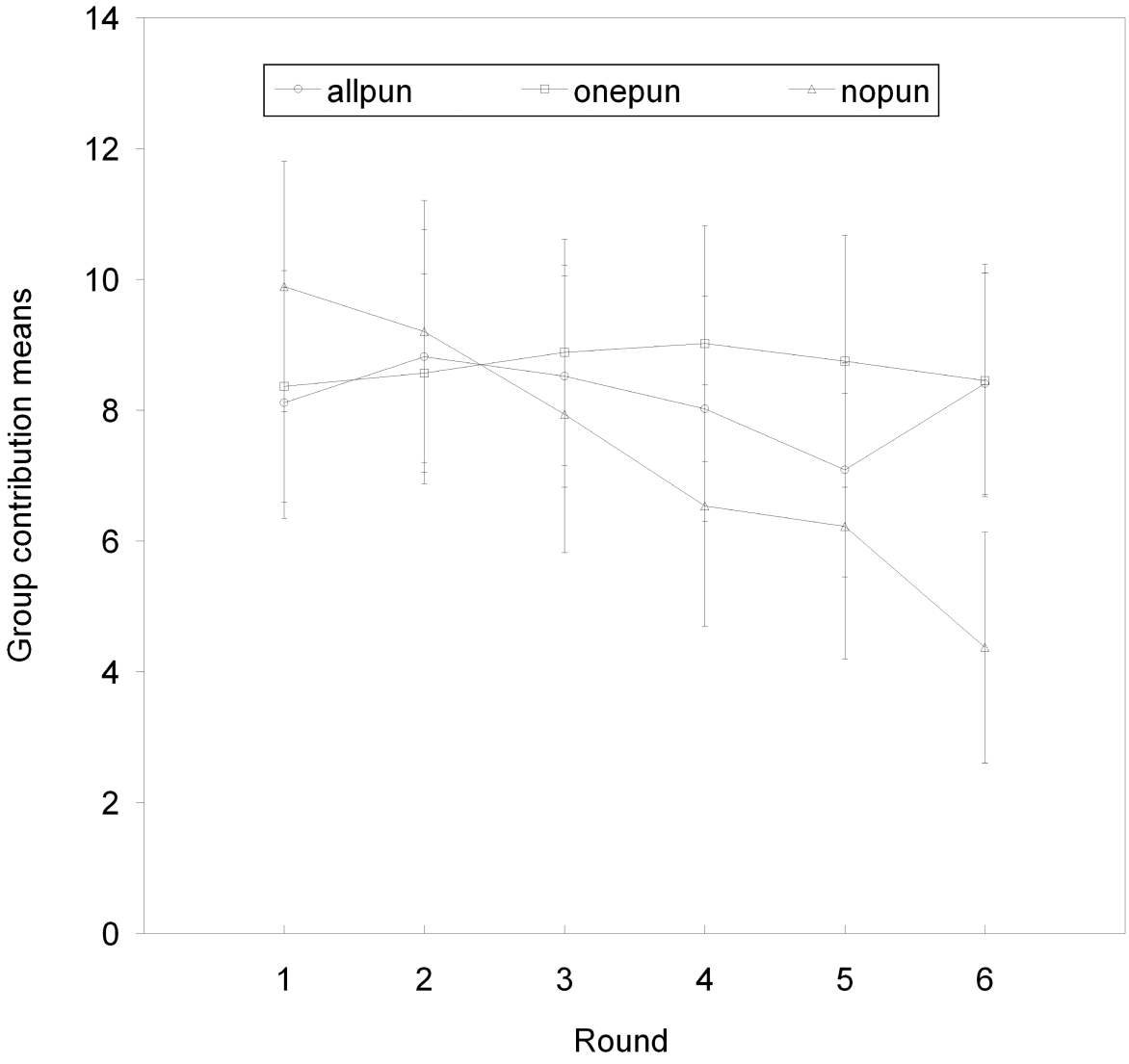
Figure legends

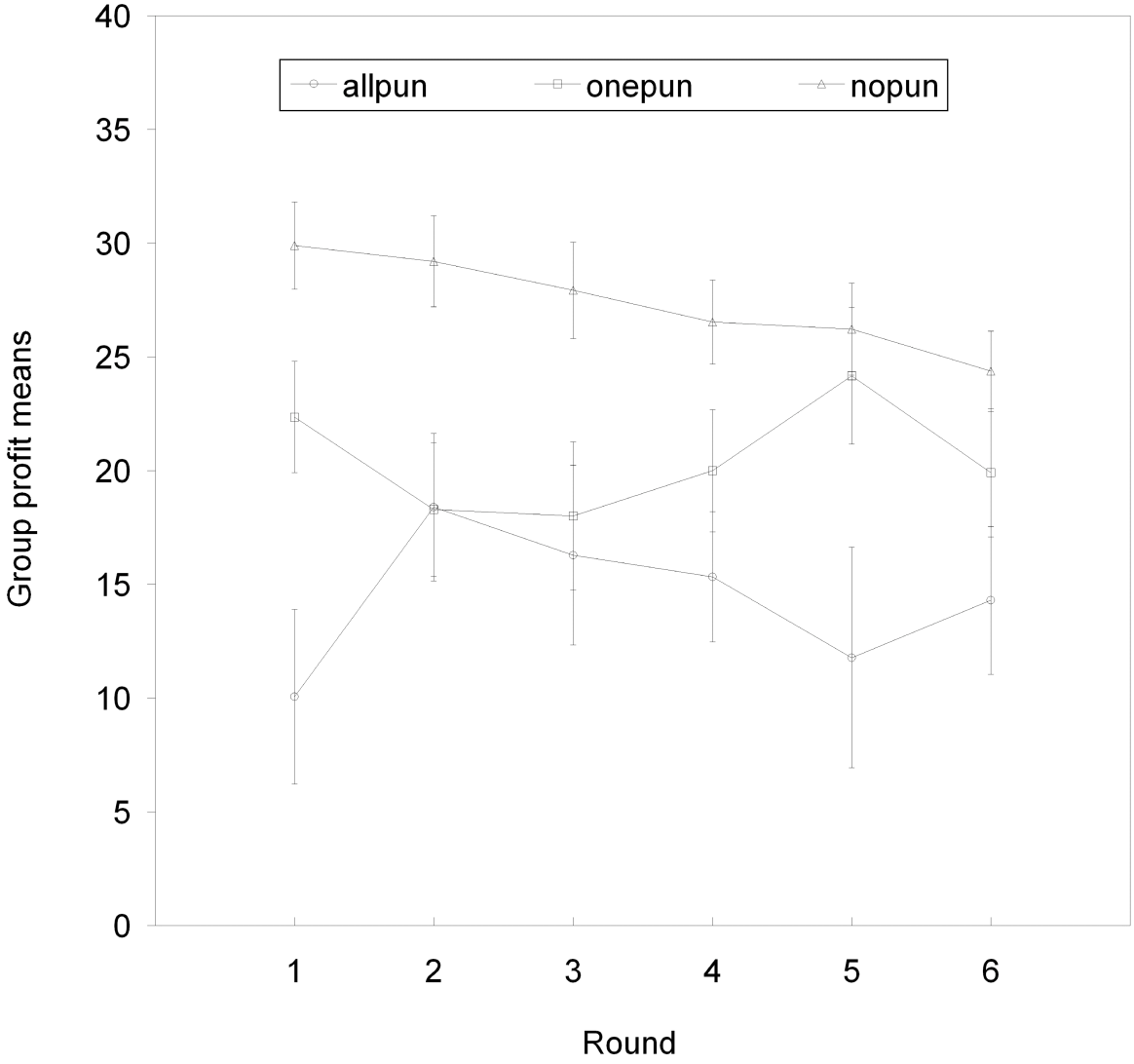
Fig.1. Mean contributions of MUs to the group fund by participants in segment two with 95% confidence intervals indicated by error bars.

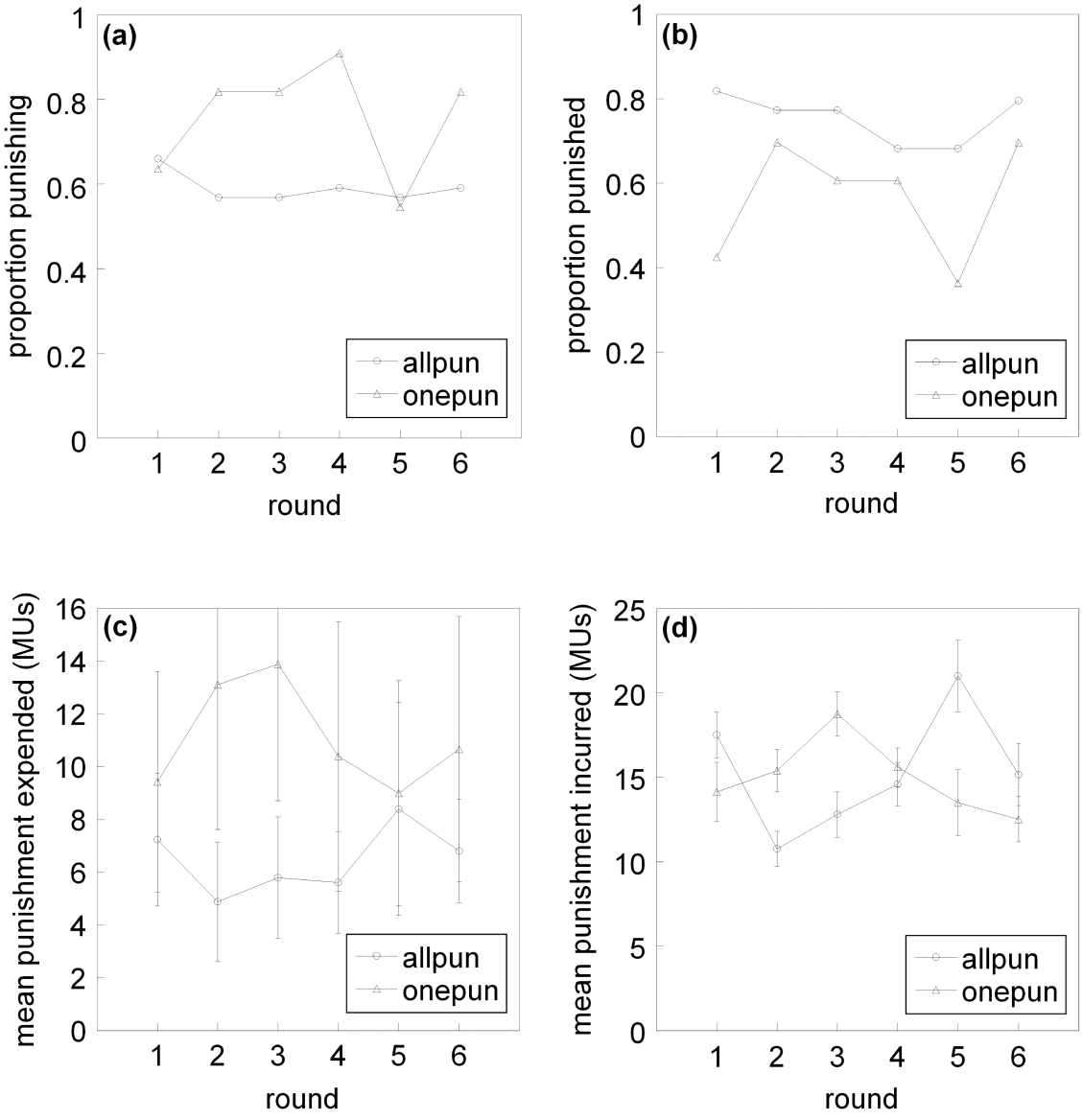
Fig.2. Mean profits (MUs) for participants in segment two with 95% confidence intervals indicated by error bars.

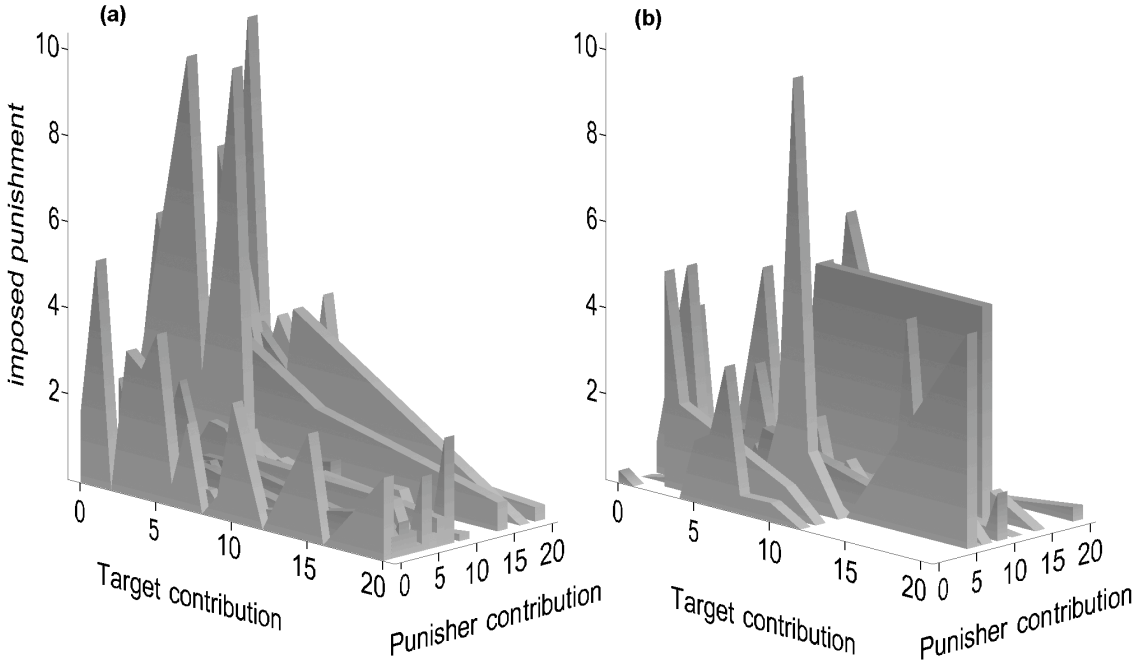
Fig. 3. There were more punishers in the one-punishment condition who punished at least once per round than in the all-punishment condition (a), although punishers in the one-punishment condition did not punish as many group members (b). One-punishment punishers did, however, expend greater resources to punish (c), resulting in a similar level of penalties being incurred within each punishment condition when considered over the six rounds in Segment Two (d).

Fig. 4. Punishers tended to apply greater deductions for values that deviated more from higher levels of possible contributions, though this effect is stronger in the all-punishment (a) than in the one-punishment condition (b).


allpun
onepun
nopun
Group contribution means
Round
0
2
4
6
8
10
12
14
1
2
3
4
5
6


allpun
onepun
nopun
Group profit means
Round
0
5
10
15
20
25
30
35
40
1
2
3
4
5
6


(a)
proportion punishing
round
allpun
onepun
(b)
proportion punished
round
allpun
onepun
(c)
mean punishment expended (MUs)
round
allpun
onepun
(d)
mean punishment incurred (MUs)
round
allpun
onepun


(a)
(b)
imposed punishment
10
8
6
4
2
0
5
10
15
20
Target contribution
Punisher contribution