Leadership in Altruistic Punishment and Reward

Simon Columbus
Amsterdam University College
University of Amsterdam
simon.hadlich@student.uva.nl

ABSTRACT

Cooperation in economic games breaks down in the absence of enforcement mechanisms. We show that in the Public Goods Game, cooperation can be sustained by altruistic punishment, but not reward. Voluntary leadership occurs frequently in both conditions, but does not affect the dynamics of contributions and enforcement. While pro-socially punishing leaders are perceived as fairer than anti-social leaders, they have a worse reputation than pro-socially rewarding leaders. Thus punishment appears to carry a reputational penalty even when it is pro-social. Contradicting predictions from indirect reciprocity theory, this provides some support for altruistic punishment as strong reciprocity.

Author Keywords

Altruistic Punishment, Altruistic Reward, Evolution of Leadership, Indirect Reciprocity, Social Coordination Hypothesis.

INTRODUCTION

The human species has an extraordinary ability to cooperate. Almost everything around us is the product of the interacting efforts of a myriad of individuals. Moreover, in contrast to other primates–which are confined to small-scale, highly related groups–humans live in large civilisations. Anonymous interactions are common in social and economic life. However, cooperation with strangers depends on the enforcement of specific social norms. Where defectors go unconstrained and free-riders can exploit public goods, cooperation breaks down [5].

Voluntary punishment and reward can sustain cooperation even when they are individually costly to the enforcer [1], under the condition that they are directed at free-riders [10]. Such *altruistic punishment* and reward occurs across a wide variety of cultures [10] [9]. Altruistic punishment and reward are thus largely established as effective enforcement mechanisms that can sustain cooperation.

SRC 2013, November 20, 2013, Amsterdam, The Netherlands.

Leadership is a second type of mechanism that can enforce cooperation [8]. Two broad theoretical frameworks have been proposed for the evolution of leadership. The Byproduct Dominance hypothesis argues that leadership and followership are "byproducts of adaptations for dominance and submission" [17]. In contrast, the Social Coordination hypothesis proposes that "leadership evolved specifically for the purpose of solving coordination problems" (ibid.). In this perspective, leadership, like punishment, would be an altruistic trait.

We test the effects of leadership in a Public Goods Game (PGG) with sequential punishment or reward. These games allow for individuals to take the charge in enforcing norms of cooperation by punishing free-riders or rewarding contributors. Policing of free-riders can be understood as a second-order collective action problem; however, leaders' options are much more ambiguous in this setting than in coordination games, where group and individual interests align. Furthermore, we consider differences in the effectiveness of punishment and reward as enforcement mechanisms. Finally, using hypothetical scenarios we test the effects of punishment and reward on the reputation of leaders.

METHODS

96 subjects (62 female) participated in a PGG with real monetary stakes and one of two treatment conditions, Punishment or Reward. In both conditions, groups of $n = 4$ members played six rounds of the PGG. Identifiers for the players were randomised at the beginning of each round such that reputation building over multiple rounds was not possible. At the beginning of each round, each of the n players in a group received an endowment of $y = 20$ tokens. During the Contribution stage, each player could decide whether to make an investment g_i of part or all of their endowment in a common project ($0 \leq g_i \leq 20$). These contributions were made simultaneously. The total investment in the common project was then multiplied by a factor $na = 1.6$ and paid out in equal parts to all players.

At the end of the Contribution stage, the full matrix of contributions was revealed to all players. During the following Enforcement round, each player could expend part of their endowment to punish or reward another player. The ratio was set such that player i has to spend $p_i = 1$ token to impose a punishment of $P^j = 3$ tokens on any player j, and similarly for reward.

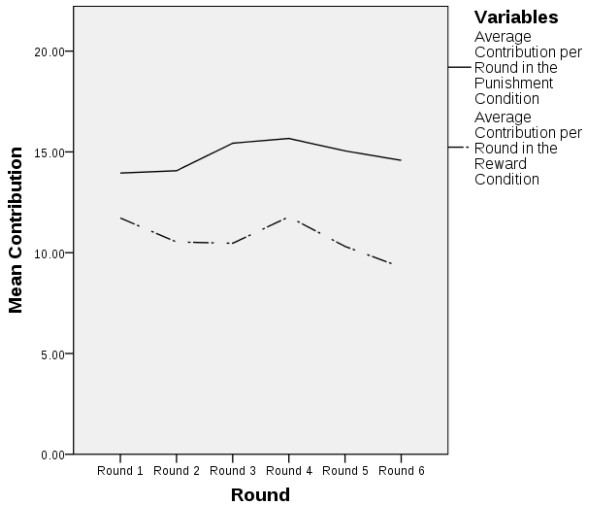

Figure 1. Changes over rounds in the average contribution to the public good.

The Enforcement round was implemented sequentially, following the design for sequential games from Gillet et al. [7]. At the start of the stage, there was a three-minute window during which the first player to move and make their punishment allocation revealed their choices to the other players. This window closed as soon as the first player made their move, and all other players then made their decisions simultaneously. If no player chose to make their allocation during the sequential punishment window, the stage would move on to the simultaneous phase. After all players had made their allocation, each player was informed about the punishment they received and their total pay-off from that round.

The total pay-off Π for player i from the PGG was their average pay-off per round t [i.e. $\Pi_i = (\sum_{t=1}^{6} \pi_i^t)/6$]. Tokens were exchanged for Euros at the end of the experiment at a rate of $3.2 : 1$, such that in the Punishment condition the earnings per player in a round with full cooperation and no punishment equal €10. For the Reward condition, tokens were exchanged for Euros at a rate of $12.2 : 1$. At this exchange rate, the earnings per player in a round with full cooperation and full reward to the player equal €10. The experiment was programmed and conducted with the software z-Tree [6].

Leaders' Reputation

Since reputation-building was not possible in the Public Goods Game due to the implementation of anonymity, the effects of punishment and reward on leaders' reputation were assessed via four hypothetical scenarios. The participants were confronted with prompts describing Pro-social (rewarding high contributors resp. punishing low contributors) and Anti-social (rewarding low contributors resp. punishing high contributors) leaders. For each hypothetical scenario, the participants were asked to rate the leader on four dimensions, 'Dominance', 'Altruism', 'Being a Good Person', and 'Fairness'. Each prompt was to be answered on a 5-option Likert scale (*1 = completely agree to 5 = completely disagree*).

RESULTS

Punishment and Reward conditions differed in total contributions and earnings, but not in the amount of points spent on enforcement–either by leaders or by all group members–nor in the distribution of leadership ($\alpha = .05$). Further analysis shows that the conditions initially did not differ significantly in contribution levels (independent-samples Mann-Whitney test, $U = 88.5, p = .65$, median reward $= 11$, median punishment $= 14$ at group level; and $U = 1,150, p = .145$, median reward $= 12$, median punishment $= 14$ at individual level). Figure 1 illustrates that average contributions to the common good increased in the Punishment condition, but fell off in the Reward condition; i.e., the trends of both conditions diverged over multiple iterations.

First-moving occurred during each round of the experiment, and always after only a fraction of the waiting period. Most players moved first at least once (Punishment $= 78\%$; Reward $= 78\%$). The median number of rounds any player chose to move first was one, and monopolisation of leadership by any one player was rare. Thus in global terms, leadership did not differ across conditions. Moreover, expenditure on enforcement did not differ significantly between leaders and non-leaders (Wilcoxon signed-rank test, $\alpha = .05$).

Reputational Consequences of Pro-social and Anti-social Leadership

Friedman tests for reputation differences between leadership styles were significant for all four variables (all $p < .001$). Pro-social Punishers and Rewarders exhibited no difference on Perceived Fairness ($p = 1.000$), and scored significantly higher than anti-social leaders ($p < .001$). Among the latter, Punishers and Rewarders did not differ significantly ($p = 1.000$).

Pro-social Rewarders were considered more altruistic than their punishing counterparts (median (both) $= 3$, $\bar{x} = 2.46$ vs. 3.07, $p = .020$). The latter were perceived as significantly more altruistic than Anti-social Punishers (median 3 vs. 4, $p = .005$), but not Anti-social Rewarders (median (both) $= 3$, $p = 1.000$). A similar pattern can be observed on the perception of being a good person. Pro-social Rewarders scored higher than Punishers (median (both) $= 3$, $\bar{x} = 2.34$ vs. 2.73, $p = .048$), and the latter were again perceived as more of a good person than Anti-social Punishers (median = 3 vs. 4, $p < .001$), but not Rewarders (median (both) $= 3$, $p = .122$). Moreover, anti-social Punishers were considered less of a good person than their Rewarding counterparts (median $= 4$ vs. 3, $p = .006$).

DISCUSSION

This study investigated a range of theories pertaining to leadership in altruistic reward and punishment, using the Public Goods Game as its underlying paradigm. First, it shows that although Reward and Punishment conditions initially do not differ significantly in contributions, they diverge over time. While contributions in the Punishment condition increase, they decrease in the Reward condition. Hence, the initial hypothesis that punishment and reward are equally effective in promoting cooperation is rejected. As participants invested a similar share of their endowments in punishing and rewarding

other players, it seems likely that the divergence of contribution levels between both conditions is due to the differential effectiveness per unit of both enforcement mechanisms (cf. [15]).

Though O'Gorman et al. [14] have proposed that altruistic punishment, but not reward, is an adaptation for the enforcement of moral norms, these observations can be explained more simply from general prospect theory. Specifically, loss-aversion is a well-established concept that describes the human tendency to value potential losses higher than potential gains [11]. If participants in the Public Goods Game exhibit loss aversion, they will react more strongly to the prospect of being punished than to potential rewards; thus making punishment the more effective enforcement mechanism.

The second focus of this study were the effects of leadership on game dynamics and the reputational benefits of reward and punishment. Leadership appears to have played a minor role in influencing game dynamics (though further analysis is in order). In terms of leader behaviour, it is at least somewhat surprising that leadership occurred in each single round of each session of the experiment; but in particular how quickly the leadership decision was regularly made. In many cases, leaders were designated within less than two seconds, and participants appeared to compete for the first-moving position. This is highly counter-intuitive given that leadership carries no formal benefits, and in particular given the threat of retaliation in the Punishment condition [3] [12].

The apparent competition for leadership in the PGG implicates several personality traits in the decision. In the Punishment condition, competition for leadership would make Dominance appear to be a more likely predictor than Altruism. In the Reward condition, the Machiavellianism hypothesis would be supported by this competition. There was, however, no significant finding among the personality correlates when attempting to either predict leadership frequency within a condition or to distinguish leader types across conditions. This may partly be due to the small sample size of the study, which is not suitable for analysis of a large number of predictors.

A second reason for competition for leadership would be reputational benefits (though such behaviour would be maladaptive given the anonymity condition of the experiment). With help of hypothetical scenarios, this study shows that pro-socially punishing leaders are considered equally fair as Pro-social Rewarders, but this fairness rating does not transfer to being perceived as a 'good person' or as altruistic. These three outcome variables are somewhat similar to Barclay's [2], who found that punishers exceeded non-punishers in ratings of trustworthiness (fairness), group focus (altruism), and being worthy of respect (being a good person). Given Barclay's positive finding for 'group focus', it is even more surprising that Pro-social Punishers were not rated more altruistic than anti-social leaders. All in all, even if punishing leaders were seen as fair, it would appear that leadership in punishment does not carry a reputational benefit.

Given that Pro-social Punishers were rated more altruistic and more of a good person than Anti-social Punishers, but not Anti-social Rewarders, and that the latter were also rated significantly better than the former, it appears that punishment itself carries a reputational cost. While prospect theory could explain why Anti-social Punishers are rated worse than Anti-social Rewarders–as high contributors would be expected to value losses due to anti-social punishment more than forfeited prospective gains due to anti-social reward to low contributors–it fails to explain the reputational penalty for Pro-social Punishers. In contrast to Barclay's [2] findings, and counter the predictions of indirect reciprocity theorists [4] [16] [13], it would then appear that punishment, even if pro-social, carries a reputational cost.

CONCLUSION

This study has found that altruistic punishment, but not reward, can sustain cooperation among an anonymous group. Though leadership occurs in both conditions, leaders do not differ significantly from followers both in terms of personality and of behaviour, and leadership does not appear to impact group dynamics. Who becomes a leader cannot be predicted by measures of altruism, dominance, Machiavellianism, and sensation-seeking. When leaders punish in the interest of the group, they are perceived as more fair than anti-social leaders, but this does not translate into reputational gains in terms of being considered altruistic or being seen as a good person.

These results confirm established findings that costly punishment is effective in sustaining cooperation. In contrast, reward appears to fail as an enforcement mechanism, despite contrary claims from earlier studies [1]. Seeing that punishment and reward occur at equal frequencies, it appears that diverging patterns in contributions result from differential effectiveness of both mechanisms in enforcing future cooperation. Prospect theory predicts that losses from punishment should be valued higher than forfeited gains from withheld reward, and might explain the greater effectiveness of punishment in enforcing cooperation.

Leadership occurred at high frequency, and there was apparent competition for moving into the leadership position. Nevertheless, leaders did not differ from followers in their expenditure on either punishment or reward. Moreover, several hypotheses on leaders' personalities that were derived from competing theories about the evolution of leadership–specifically the Social Coordination and Dominance Byproduct hypotheses–could not be confirmed. There was no apparent pattern in leaders' personalities, neither compared to followers within the same condition nor across conditions.

In hypothetical scenarios, leaders who used punishment prosocially did not earn reputational benefits. Although Pro-social Punishers were perceived as equally fair as Pro-social Rewarders (and as more fair than anti-social leaders), they scored lower on measures of altruism and being a good person, and no different from Anti-social Rewarders. It thus appears that punishment itself carries a reputational cost, even when it is used to benefit others.

Future research should focus on the dynamics of the Public Goods Game between stages, rounds, leaders, followers,

groups, and conditions. Multi-level modelling techniques appear best-suited for these research questions. Such analysis of the present data may yield further insight into the presence of leadership by example, counter-punishment, and reciprocation of reward; as well as into the effects of punishment and reward on contributions in subsequent rounds.

Furthermore, the reputational effects of enforcement and leader behaviour in the Public Goods Game should be assessed using actual leader ratings rather than hypothetical scenarios. So far, there is evidence both in favour [2] and against reputational benefits to punishment. Further research should specify these effects for leaders (as opposed to peer enforcement).

Altruistic punishment and reward, leadership, and reputation are most likely evolved adaptations that enable life in large-scale societies where cooperation with strangers is common. On their own, but even more so in conjunction with cultural institutions, they enable humanity to undertake astonishing enterprises. Understanding their dynamic interactions can inform the design of institutions today, from the governance of common pool resources to the management of businesses and the organisation of whole societies.

ROLE OF THE STUDENT

SC undertook this study as a capstone project towards a BA degree in Social Sciences at Amsterdam University College, under supervision of Prof. Dr. Mark van Vugt of VU University Amsterdam. The topic was jointly conceived by SC, MvV, and Dr. Thomas Pollet. SC independently acquired funding, devised the treatments and questionnaires, implemented the experiments in z-Tree, ran sessions, analysed data, and wrote up results.

ACKNOWLEDGMENTS

I thank Mark van Vugt and Thomas Pollet for their support and guidance, as well as Sennay Ghebreab for providing me with access to the B&TA Lab. Further thanks go to Fayette Klaassen, Luc Draisma, Martine van Dusschoten, Lukas Snoek, Jenny Rutten, Bente de Vries, Henk Nieweg, Alina Berendsen, Agnese Logina, Francesca Grandolfo, and Joris Gillet for their help at various stages of this study.

This study was sponsored by the Beta Beurs. The Beta Beurs is devised by the Brain & Technology Amsterdam (B&TA) Lab, co-funded by the Center for Creation, Content and Technology (CCCT), and subsidised by Platform Beta Techniek (PBT).

REFERENCES

1. Balliet, D., Mulder, L. B., and Van Lange, P. a. M. Reward, punishment, and cooperation: a meta-analysis. *Psychological Bulletin 137*, 4 (July 2011), 594–615.
2. Barclay, P. Reputational benefits for altruistic punishment. *Evolution and Human Behavior 27*, 5 (Sept. 2006), 325–344.
3. Denant-Boemont, L., Masclet, D., and Noussair, C. N. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory 33*, 1 (Jan. 2007), 145–167.
4. dos Santos, M., Rankin, D. J., and Wedekind, C. The evolution of punishment through reputation. *Proceedings of the Royal Society. Series B, Biological Sciences 278*, 1704 (Feb. 2011), 371–7.
5. Fehr, E., and Schmidt, K. M. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics 114*, 3 (Aug. 1999), 817–868.
6. Fischbacher, U. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*, 2 (Feb. 2007), 171–178.
7. Gillet, J., Cartwright, E., and Vugt, M. V. Selfish or servant leadership? Evolutionary predictions on leadership personalities in coordination games. *Personality and Individual Differences 51*, 3 (Aug. 2011), 231–236.
8. Güth, W., Levati, M. V., Sutter, M., and van der Heijden, E. Leadership and cooperation in public goods experiments. 2004.
9. Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., and Ziker, J. Markets, religion, community size, and the evolution of fairness and punishment. *Science (New York, N.Y.) 327*, 5972 (Mar. 2010), 1480–4.
10. Herrmann, B., Thöni, C., and Gächter, S. Antisocial punishment across societies. *Science 319*, 5868 (Mar. 2008), 1362–7.
11. Kahneman, D., and Tversky, A. Choices, Values, and Frames. *American Psychologist 39*, 4 (1984), 341–350.
12. Nikiforakis, N. Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics 92*, 1-2 (Feb. 2008), 91–112.
13. Nowak, M. A., and Highfield, R. *SuperCooperators: Altruism, Evolution, and Why We Need Each Other to Succeed*. Free Press, New York, NY, 2011.
14. O'Gorman, R., Wilson, D. S., and Miller, R. R. Altruistic punishing and helping differ in sensitivity to relatedness, friendship, and future interactions. *Evolution and Human Behavior 26*, 5 (Sept. 2005), 375–387.
15. Sefton, M., Shupp, R., and Walker, J. M. The Effect of Rewards and Sanctions in Provision of Public Goods. *Economic Inquiry 45*, 4 (Oct. 2005), 671–690.
16. Sigmund, K., Hauert, C., and Nowak, M. A. Reward and punishment. *Proceedings of the National Academy of Sciences of the United States of America 98*, 19 (Sept. 2001), 10757–62.
17. van Vugt, M. Evolutionary Origins of Leadership and Followership. *Personality and Social Psychology Review 10*, 4 (2006), 354–71.