Mapping the punishment strategy space

Scott Claessens[*1], Quentin D. Atkinson[1], & Nichola Raihani[1,2]

[1] School of Psychology, University of Auckland, Auckland, New Zealand

[2] Department of Experimental Psychology, University College London, London, United Kingdom

* Correspondence concerning this article should be addressed to Scott Claessens, Level 2, Building 302, 23 Symonds Street, Auckland, New Zealand. E-mail: scott.claessens@gmail.com

This working paper has not yet been peer-reviewed.

Abstract

x

Keywords: x

Word count: xxxx words

Mapping the punishment strategy space

Introduction

Humans cooperate on a scale unparalleled in the animal kingdom. One mechanism thought to sustain this level of cooperation is costly punishment, whereby individuals harm others at a personal cost[1], ostensibly encouraging cooperative behaviour from the target (or bystanders[2–4]) in the future. Punishment therefore offers a route to maintaining or increasing cooperation by changing the payoff structure of social interactions such that it no longer pays to cheat or exploit social partners[1,5].

In humans, many studies of punishment have been carried out in laboratory settings using economic games[6–15]. In these games, participants are usually given a sum of money that they can use to invest in collective action or to help others. Alternatively, participants can 'cheat' by keeping the money for themselves or by exploiting the contributions of others. Punishment is introduced into such games by giving participants the option to pay a small 'fee' to impose a greater 'fine' on their co-players. Several lines of experimental evidence indicate that people use this punishment option[12], that they enjoy punishing[16], and that they frequently, though not always[17], punish cheating or exploitative co-players[10,11].

Evidence from these experiments suggests that the threat of costly punishment plays an important role in promoting human cooperation. People tend to cooperate more in games where punishment is possible compared to those where it is not[6,7,15]. The effect that the threat of punishment has on cooperation is also evident in the higher contributions typically observed in the Ultimatum Game (where punishment is possible) compared to the structurally-similar Dictator Game (where it is not)[18]. This typical cooperation-enhancing effect of punishment has also been observed across societies[7], leading some to suggest that costly punishment has played a key role in the cultural evolution of cooperation in humans[19–22].

Nevertheless, it remains unclear whether individuals playing economic games use punishment as a behaviour-change tool to enforce cooperation or as a means to achieve other ends. Some have argued that punishment is primarily used to shape future behaviour, either to deter personal harm[3,9,23] or to uphold normative standards of cooperative behaviour[20,21,24–27]. But while the *threat* of punishment can have a cooperation-enhancing effect, the *enactment* of this punishment seldom changes the future behaviour of deviant targets[15]. This calls into question whether punishment primarily operates as a behaviour-change tool or whether it is used to achieve other goals.

Beyond behaviour-shaping concerns, there are a host of other reasons that people may want to punish in economic games. Punishers might be motivated by a desire for retribution rather than deterrence, punishing in proportion to the amount of harm that was personally caused[28]. Punishment might be driven by concerns about relative payoffs, such as disadvantageous inequity aversion (i.e., avoiding having less than others[15,29]) and/or general egalitarian preferences (i.e., wanting all participants to receive the same payoffs[30]). Such concerns about relative payoffs may be activated when participants earn less than cheats in economic games or when there are income disparities in these settings. People might also use punishment for competitive purposes, seeking advantageous inequity for themselves (i.e., having more than others) and/or improving their relative position[15].

Common economic game designs have been unable to tease apart these different motives for punishment because participants who interact with cheats in these games experience both losses *and* lower relative payoffs. The typical 1:3 fee-to-fine ratio of punishment in economic games compounds this issue. With this setup, people can simultaneously use punishment to reciprocate losses, to deter others from cheating, and to reduce or reverse disparities in payoffs between themselves and targets. To add to this complexity, it is evident that people use punishment in seemingly disparate ways: punishing when no behaviour change is possible, such as in one-shot games[12,29,31,32] or in games where the target never learns about the punishment[33]; punishing those who did not cheat or who

over-contributed to collective action (antisocial punishment[17,34]); punishing in scenarios where they were not personally harmed (third-party punishment[35]); and punishing in scenarios where disparities in payoffs did not arise from participants' actions[30,36,37].

The general conclusion from this research is that there is no one unifying function of costly punishment in humans. Instead, punishment should be thought of as a flexible behavioural tool that serves a variety of functions that are not mutually exclusive[15]. Due to its multipurpose nature, we should therefore expect variation in punishment strategies in the population, much like the observed variation in social learning strategies[38]. Some individuals may use punishment as a behaviour shaping tool, for example, while others may use it to reduce or reverse payoff differentials.

This insight raises several underexplored questions. First, which punishment strategies are more frequent in human populations? Second, what traits predict adherence to a particular strategy? Previous work has reported that personality is related to cooperative behaviour[39] and demographics, political ideology, and religiosity are related to punitive behaviour[40], but no research has related these variables to specific punishment strategies. Third, do people have insight into their own punishment strategy? Previous work has argued that people are often unaware of the underlying function of their punitive behaviour, yet they feel compelled to enact it anyway[28,41].

Here, we aim to delineate the many possible punishment strategies by asking whether people punish in a manner consistent with a specific strategy and, if so, what other characteristics (personality, social preferences, political orientation) predict the use of different punishment strategies. Table 1 summarises the potential functions for costly punishment in economic games and the behavioural strategies they predict. Note that we do not include reputational functions of punishment in this table, such as signalling trustworthiness[4,42–45], because our focus is on punishment strategies in anonymous economic games without reputational incentives (but see ref[46]).

Table 1

xxx

Function	Behavioural strategy	No DI 1 70-10		No DI 2 70-30		No DI 3 (Computer) 70-30		No DI 4 (1:1 Fee-Fine) 70-30		DI 70-50		3PP 70-70 [100]	
		Take 50-30	No take 70-10	Take 50-50	No take 70-30	Take 50-50	No take 70-30	Take 50-50	No take 70-30	Take 50-70	No take 70-50	Take 50-90 [100]	No take 70-70 [100]
Deterrent	Punish to deter another who has harmed you from harming you again in the future	✓	X	✓	X	X	X	✓	X	✓	X	X	X
Norm-enforcing	Punish to enforce a shared anti-harm norm and encourage future norm compliance, even amongst third parties	✓	X	✓	X	X	X	✓	X	✓	X	✓	X
Retributive	Punish if doing so harms another who has harmed you	✓	X	✓	X	✓	X	✓	X	✓	X	X	X
Avoid DI	Punish if doing so avoids disadvantageous inequity for self	X	X	X	X	X	X	X	X	✓	X	X	X
Egalitarian	Punish if doing so makes payoffs for all more equal	X	X	X	X	X	X	X	X	✓	X	✓	X
Seek AI	Punish if doing so produces advantageous inequity for self	X	X	✓	X	✓	X	X	X	X	X	X	X
Competitive	Punish if doing so improves your relative position	✓	✓	✓	✓	✓	✓	X	X	✓	✓	✓	✓
Antisocial	Punish exclusively those who do not cause harm	X	✓	X	✓	X	✓	X	✓	X	✓	X	✓
Never punish	Never punish others	X	X	X	X	X	X	X	X	X	X	X	X

Building on previous designs[29,31,47,48], we employ a suite of one-shot economic games where individuals are given the opportunity to punish targets at a personal cost. We carefully designed this suite of games to tease apart the proposed punishment strategies in Table 1, such that each strategy predicts a different pattern of behaviour across all the games. We use the resulting behavioural patterns to discern which punishment strategy participants are employing. We then combine these behavioural patterns with data on demographics, personality, political ideology, religiosity, and self-reported strategy usage.

Results

x

Discussion

x

Methods

x

Acknowledgements

x

Author Contributions

x

Competing Interests

The authors declare no competing interests.

Data Availability

All data used in this study are publicly available on GitHub: https://github.com/ScottClaessens/punishStrategies

Code Availability

All code to reproduce the analyses in this study are publicly available on GitHub: https://github.com/ScottClaessens/punishStrategies

References

1. Clutton-Brock, T. H. & Parker, G. A. Punishment in animal societies. *Nature* **373**, 209–216 (1995).

2. dos Santos, M., Rankin, D. J. & Wedekind, C. The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences* **278**, 371–377 (2011).

3. dos Santos, M., Rankin, D. J. & Wedekind, C. Human cooperation based on punishment reputation. *Evolution* **67**, 2446–2450 (2013).

4. Raihani, N. J. & Bshary, R. The reputation of punishers. *Trends in Ecology & Evolution* **30**, 98–103 (2015).

5. Ostrom, E. *Governing the commons: The evolution of institutions for collective action*. (Cambridge University Press, 1990).

6. Balliet, D., Mulder, L. B. & Van Lange, P. A. M. Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin* **137**, 594–615 (2011).

7. Balliet, D. & Lange, P. A. M. V. Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspectives on Psychological Science* **8**, 363–379 (2013).

8. Chaudhuri, A. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics* **14**, 47–83 (2011).

9. Dreber, A., Rand, D. G., Fudenberg, D. & Nowak, M. A. Winners don't punish. *Nature* **452**, 348–351 (2008).

10. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *American Economic Review* **90**, 980–994 (2000).

11. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).

12. Henrich, J. *et al.* Costly punishment across human societies. *Science* **312**, 1767–1770 (2006).

13. Nikiforakis, N. & Normann, H.-T. A comparative statics analysis of punishment in public-good experiments. *Experimental Economics* **11**, 358–369 (2008).

14. Raihani, N. J., Thornton, A. & Bshary, R. Punishment and cooperation in nature. *Trends in Ecology & Evolution* **27**, 288–295 (2012).

15. Raihani, N. J. & Bshary, R. Punishment: One tool, many uses. *Evolutionary Human Sciences* **1**, e12 (2019).

16. de Quervain, D. J.-F. *et al.* The neural basis of altruistic punishment. *Science* **305**, 1254–1258 (2004).

17. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367 (2008).

18. Camerer, C. F. *Behavioral game theory: Experiments in strategic interaction.* (Russell Sage Foundation, 2003).

19. Bowles, S. & Gintis, H. *A cooperative species: Human reciprocity and its evolution.* (Princeton University Press, 2013).

20. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences* **100**, 3531–3535 (2003).

21. Chudek, M. & Henrich, J. Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences* **15**, 218–226 (2011).

22. Henrich, J. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter.* (Princeton University Press, 2017).

23. Delton, A. W. & Krasnow, M. M. The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior* **38**, 734–743 (2017).

24. Fehr, E. & Schurtenberger, I. Normative foundations of human cooperation. *Nature Human Behaviour* **2**, 458–468 (2018).

25. Mathew, S. & Boyd, R. Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences* **108**, 11375–11380 (2011).

26. Mathew, S. & Boyd, R. The cost of cowardice: Punitive sentiments towards free riders in Turkana raids. *Evolution and Human Behavior* **35**, 58–64 (2014).

27. Richerson, P. *et al.* Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences* **39**, e30 (2016).

28. Carlsmith, K. M., Darley, J. M. & Robinson, P. H. Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* **83**, 284–299 (2002).

29. Raihani, N. J. & McAuliffe, K. Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters* **8**, 802–804 (2012).

30. Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R. & Smirnov, O. Egalitarian motives in humans. *Nature* **446**, 794–796 (2007).

31. Bone, J. E. & Raihani, N. J. Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior* **36**, 323–330 (2015).

32. Walker, J. M. & Halloran, M. A. Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics* **7**, 235–247 (2004).

33. Crockett, M. J., Özdemir, Y. & Fehr, E. The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General* **143**, 2279–2286 (2014).

34. Sylwester, K., Herrmann, B. & Bryson, J. J. Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics* **6**, 167–188 (2013).

35. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evolution and Human Behavior* **25**, 63–87 (2004).

36. Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R. & Smirnov, O. The role of egalitarian motives in altruistic punishment. *Economics Letters* **102**, 192–194 (2009).

37. Fowler, J. H., Johnson, T. & Smirnov, O. Egalitarian motive and altruistic punishment. *Nature* **433**, E1 (2005).

38. Molleman, L., Van den Berg, P. & Weissing, F. J. Consistent individual differences in human social learning strategies. *Nature Communications* **5**, 3570 (2014).

39. Thielmann, I., Spadaro, G. & Balliet, D. Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin* **146**, 30–90 (2020).

40. Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B. & Skitka, L. J. Moral punishment in everyday life. *Personality and Social Psychology Bulletin* **44**, 1697–1711 (2018).

41. Nisbett, R. E. & Wilson, T. D. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* **84**, 231–259 (1977).

42. Barclay, P. Reputational benefits for altruistic punishment. *Evolution and Human Behavior* **27**, 325–344 (2006).

43. Batistoni, T., Barclay, P. & Raihani, N. J. Third-party punishers do not compete to be chosen as partners in an experimental game. *Proceedings of the Royal Society B* **289**, 20211773 (2022).

44. Jordan, J. J., Hoffman, M., Bloom, P. & Rand, D. G. Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476 (2016).

45. Jordan, J. J. & Rand, D. G. Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology* **421**, 189–202 (2017).

46. Jordan, J. J. & Rand, D. G. Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology* **118**, 57–88 (2020).

47. Deutchman, P., Bračič, M., Raihani, N. J. & McAuliffe, K. Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. *Evolution and Human Behavior* **42**, 12–20 (2021).

48. Marczyk, J. Human punishment is not primarily motivated by inequality. *PLoS ONE* **12**, e0171298 (2017).

Supplementary Information

Mapping the punishment strategy space

Scott Claessens[1], Quentin D. Atkinson[1], Nichola Raihani[1,2]

[1] School of Psychology, University of Auckland, Auckland, New Zealand

[2] Department of Experimental Psychology, University College London, London, United Kingdom

Supplementary Figures

Supplementary Tables

Supplementary References