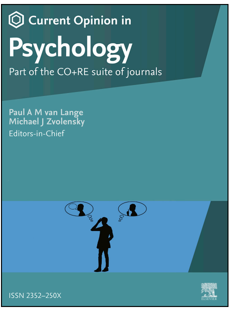


Reward, Punishment, and Prosocial Behavior: Recent Developments and Implications

Junhui Wu, Shenghua Luan, Nichola Raihani

PII: S2352-250X(21)00172-X

DOI: https://doi.org/10.1016/j.copsyc.2021.09.003

Reference: COPSYC 1250

To appear in: *Current Opinion in Psychology*

Received Date: 29 July 2021

Revised Date: 31 August 2021

Accepted Date: 2 September 2021

Please cite this article as: Wu J, Luan S, Raihani N, Reward, Punishment, and Prosocial Behavior: Recent Developments and Implications, *Current Opinion in Psychology*, https://doi.org/10.1016/j.copsyc.2021.09.003.

Reward, Punishment, and Prosocial Behavior: Recent Developments and Implications

Junhui Wu[1,2*], Shenghua Luan[1,2], Nichola Raihani[3]

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China

[3] Department of Experimental Psychology, University College London, London WC1H 0AP, UK

Author Note

* Correspondence concerning this article should be addressed to Junhui Wu, Institute of Psychology, Chinese Academy of Sciences, No. 16 Lincui Road, Chaoyang District, Beijing 100101, China. Email: wujunhui@psych.ac.cn.

Abstract (115 words)

Reward and punishment change the payoff structures of social interactions and therefore can potentially play a role in promoting prosocial behavior. Yet, there are boundary conditions for them to be effective. We review recent work that addresses the conditions under which rewards and punishment can enhance prosocial behavior, the proximate and ultimate mechanisms for individuals' rewarding and punishing decisions, and the reputational and behavioral consequences of reward and punishment under noise. The reviewed evidence points to the importance of more field research on how reward and punishment can promote prosocial behavior in real-world settings. We also highlight the need to integrate different methodologies to better examine the effects of reward and punishment on prosocial behavior.

Keywords: reward, punishment, sanctions, reputational benefits, prosocial behavior

Highlights (optional)

- Both reward and punishment can promote prosocial behavior but are costly to enact.
- Reward is less costly than punishment to implement when prosociality is rare.
- Decisions to reward and punish are driven by different emotions and motives.
- How reward and punishment operate under noise is important to address.

Reward, Punishment, and Prosocial Behavior: Recent Developments and Implications

1. Introduction

Prosocial behavior refers to a broad category of behaviors (e.g., helping, volunteering, charitable donation, and cooperative behavior) that are generally beneficial to others but often at a personal cost to the actor [1]. Prosocial behavior is critical for interpersonal relationships, groups, and societies at large to function well. For instance, engaging in prosocial behavior can enhance the actor's well-being [2,3], can improve employees' performance in organizational settings [4], and is critical to solve global social dilemmas, such as climate change and mitigating pandemics [5,6]. Researchers across different disciplines have examined the antecedents of prosocial behavior. In particular, reward and punishment have been identified as two major structural solutions that change the payoffs of different courses of actions and thus can promote prosocial behavior [7,8].

Reward and punishment are both temporarily costly actions that result in an immediate benefit or cost for the rewarded or punished target, respectively. Reward is typically targeted at prosocial actors, whereas punishment is more often levelled at free riders in social interactions [8–10]. Early research focused mainly on whether reward and punishment can increase prosocial behavior, often in laboratory experiments using social dilemma paradigms (e.g., public goods game; see Figure 1 for illustrations) [11,12], and a large-scale meta-analysis indicated that reward and punishment have similar-sized positive effects on prosocial behavior [7]. Yet, a closer examination of existing studies shows mixed evidence [10,13], suggesting that there might be boundary conditions for reward and punishment to be effective.

In this review, we summarize recent developments pertaining to three major questions (see Figure 2 for an overview): (a) do reward and punishment promote prosocial behavior and, if so, when? (b) why and when are people willing to reward or punish? (c) what are the reputational and behavioral consequences of reward and punishment under noise? We end by discussing the implications of these developments for future research.

2. When do reward and punishment promote prosocial behavior?

Reward and punishment are both behaviors that require the actor to pay a short-term cost, but they differ in the consequences for the target: reward generates immediate payoffs for the target, whereas punishment does the opposite. Hence, punishing free riders typically reduces collective payoffs and thus can often be less efficient than simply withholding help from free riders [9]. In addition, punishment can sometimes prompt retaliation rather than prosocial behavior in public goods games, thereby lowering contributions to public goods [10,14]. This negative consequence is particularly likely when punishment behaviors are believed to stem from malign motives [15] or perceived to be less legitimate [16]. For instance, punishment enacted by an uninvolved bystander (third-party punishment) or through a democratic process of majority vote (democratic punishment) are both typically perceived as more legitimate than direct punishment by the targets' interaction partner(s), and may therefore be more likely to induce targets' prosocial behavior (for a review, see [10]).

Some studies have found that reward can be more likely than punishment to promote prosocial behavior, such as inducing more contributions to public goods [17]. However, both reward and punishment can also have negative effects, such as crowding out individuals' intrinsic motivation to act prosocially [18,19]. Moreover, although third-party reward (i.e., an uninvolved bystander rewards a prosocial actor), also known as indirect reciprocity, can theoretically maintain prosocial behavior among unrelated strangers, there has been no consensus on what social rules that people use to assess others' actions (e.g., whether helping a free rider is good and deserves to be rewarded) can best promote prosociality through indirect reciprocity [20]. In order for indirect reciprocity to sustain prosocial behavior, theoretical models require that individuals should discriminate between justified defection and unjustified defection, such that an actor who refuses to help a free rider is perceived as good and gets rewarded [21]. Yet, recent evidence suggests that people evaluated justified defectors as neither good nor bad [22], which deviates from theoretical predictions. As a result, it is unclear whether in the real world, such social rules are frequently used and work effectively to sustain prosocial behavior through indirect reciprocity.

Notably, punishment and reward may be most effective when they are used in tandem, rather than separately. In particular, theoretical evidence from evolutionary models shows that reward is essential to establish prosociality when prosociality is rare in the group, whereas punishment is instrumental to maintain prosociality when the number of prosocial actors exceeds a certain threshold [5,23].

3. Why and when are people willing to reward or punish?

Here we ask what proximate and ultimate mechanisms underpin individuals' tendency to reward or punish others in social interactions (see Figure 2 for an overview). One general finding is that when given the choice, people typically prefer to reward prosocial actors (or to perform other positive actions such as compensating the victim) than to punish norm violators [24–26]. Rewarding decisions by third-party observers may be prompted by the positive affect they experience when they learn about others' prosocial behavior, and this positive affect may prompt their decisions to reward those prosocial actors [27]. People are also more prone to reward prosocial actors who are authentically motivated to care about others' welfare and are perceived as genuinely moral, such as when prosocial acts are targeted at lower-power recipients (see Figure 2) [28]. Notably, individuals who reward prosocial actors or compensate the victim are more positively evaluated by third-party observers and are also more likely to be chosen as potential interaction partners than punishers [29–31]. Such opportunities for reputational benefits may help illuminate the ultimate (evolutionary) explanations for why people are willing to pay to reward prosocial actors.

In contrast to rewarding decisions, more research has focused on the motives prompting punishment decisions. Evidence suggests that people willingly pay to punish norm violators in experimental settings and such punishment is subjectively rewarding [32]. Negative emotions, particularly anger and moral outrage, seem to reliably predict punishment decisions [33–35], including third-party punishment [36]. Indeed, introducing a time delay between norm violations and punishment decisions has been found to reduce punishment behavior [37], which is consistent with the idea that punishment is prompted by negative emotions. Similarly, evidence suggests that people also punish less often and more mildly when they make punishment decisions before (instead of immediately after) the occurrence of others' norm violations [38]. But not all punishment is motivated by anger: For example, third-party punishment can also be motivated by compassion toward the victims [39], as well as punishers' incidental feelings of gratitude induced by recalling past events (e.g., recalling a time that they were grateful) [40].

Some recent studies also suggest that people tend to attune their punishment decisions to the potential benefits of changing the target's behavior and the costs of potential retaliation [34]. For instance, people are more likely to engage in third-party punishment to deter the target from acting against their own interests when they expect future interactions with the target [41,42]. People are also more likely to punish when they value the victims' welfare and perceive that the harm to the victims has produced a net cost to themselves (i.e., the punisher has a stake in the victims' welfare), for example, when the victims are their siblings and close friends rather than their acquaintances [33,43]. In addition, people with higher power or social status, who are less likely to be retaliated against, are expected to punish [44] and are indeed more willing to punish norm violators [34,45].

Finally, individuals' group membership can affect when and how they choose to reward or punish others. As third-party observers, people tend to punish selfish behaviors committed by outgroup members more harshly than similar behaviors committed by ingroup members, which helps protect their ingroup members from exploitation or harm by the outgroup in the future [41,46]. Also, during intergroup conflicts, people are often more willing to punish free riders and reward cooperators within their group at some personal costs, because this enhances within-group cooperation, thereby making group success more likely (see Figure 2) [47].

The ultimate causes for punitive sentiment to be under positive selection also include the opportunities for reputational benefits (particularly for third-party punishers) [48], but some punishment may also be favored because it improves the punisher's payoffs or status relative to the payoffs or status of the target [10,49].

4. Reward and punishment under noise

Experimental research often assumes perfect monitoring, such that everyone can observe everyone else’s actual behavior and can reward or punish appropriately [11,12]. Yet, real-life social interactions often contain “noise”—unintended errors that cause discrepancies between intended outcomes and actual outcomes [50]. Such noise may cause imperfect monitoring and false reputations (e.g., prosocial actors are perceived as free riders), and mislead people to reward prosocial actors who are actually free riders and to punish free riders who are actually prosocial actors. Inappropriate rewarding and punishing behaviors caused by noise may eventually undermine prosocial behavior and affect the reputations of rewarders and punishers. For instance, studies exploring how leaders’ reputations are affected by noise-induced mistakes in punishing or rewarding others found that mistaken punishment damages leaders’ reputation, whereas mistaken reward does not [51]. This may occur because punishment is a harmful act and is therefore judged more negatively than reward when it is applied inappropriately. Moreover, noise may hinder the positive effects of reward and punishment on prosocial behavior. For instance, when there is a higher degree of noise, people tend to increase their punishment expenditures, but punishment cannot maintain a high contribution level and even harms group payoffs in such situations due to the possibility of mistakenly punishing high contributors [52]. Other evidence from evolutionary models on institutional reward and punishment suggests that for intermediate and high levels of noise, reward performs best in eliciting higher contribution levels and group welfare, whereas punishment fails to maintain a high contribution level and thereby reduces group welfare [53].

Undoubtedly, to better understand how reward and punishment can be used to promote prosocial behavior in real-life situations, it is important for future research to pay more attention to the effects of reward and punishment on prosocial behavior under noise, which have been relatively understudied (see Figure 2). It is also important to note that people in real-life situations can also learn about others' behavior through gossip when they cannot directly observe these others' behavior. Gossip may be best able to overcome the problem of noise when it comes from multiple independent sources [54].

5. Implications and conclusions

Existing research on reward and punishment, largely relying on evolutionary models and laboratory experiments, has suggested that reward and punishment are generally effective means to promote prosocial behavior. Yet, peer punishment seems to work less efficiently than reward and other forms of punishment, such as third-party punishment and democratic punishment [5,55–57]. Notably, punishments enacted in the laboratory often differ from those observed in real-life social interactions (e.g., [43,58]), because people in real-life situations can often intervene in multiple ways, including through direct physical or verbal confrontation, and indirect reputation-based strategies, such as social avoidance and gossip [34]. Both field and laboratory studies have shown that gossip and social image concerns can promote prosocial behavior more efficiently than punishment [59,60]. It is possible that people may first gossip about others' norm violations and then coordinate their punishment behaviors if gossip alone does not work. In addition, how reward works compared to indirect strategies (e.g., social avoidance and gossip) has been relatively understudied. Future research can use multi-trial tasks to examine the dynamic changes in the uses of reward, punishment, and indirect reputation-based strategies, and how they can be combined to more efficiently promote and sustain prosocial behavior.

Another observation from this selective review is that there has been a plethora of research using evolutionary models to investigate the optimal conditions for reward and punishment to promote and sustain prosocial behavior [5,23,53,61]. However, whether results from evolutionary models can accurately reflect individuals' behavioral patterns in experiments and real-life interactions remains unknown. For example, although modeling results suggest that the best strategy to solve social dilemmas is to use reward first and then switch to punishment when the number of prosocial actors reaches a certain threshold [5,23], this prediction has not yet been tested in empirical studies. To provide more useful insights for policy makers, future research needs to integrate modeling approaches with behavioral and field studies to generate more ecologically valid and robust findings with regard to the effectiveness of different structural solutions.

To conclude, despite the overall effectiveness of reward and punishment in promoting prosocial behavior, we should be aware of the boundary conditions for them to work effectively without harming collective welfare. In addition, decisions to reward and punish are driven by different emotions and motives, which can provide useful insights into how to encourage the provision of reward and punishment systems to enhance prosocial behavior. Notably, more field research is needed on how reward and punishment, compared to indirect reputation-based strategies such as social avoidance and gossip, promote prosocial behavior in "noisy" real-world settings.

Funding

This work was supported by funding from the National Natural Science Foundation of China (71901028) to Junhui Wu, funding from Chinese Academy of Sciences (Y8CX153006) to Shenghua Luan, and funding from Royal Society University Research Fellowship and Leverhulme Trust to Nichola Raihani.

Conflict of interest statement

Nothing declared.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

** of outstanding interest

1. Penner LA, Dovidio JF, Piliavin JA, Schroeder DA: **Prosocial behavior: Multilevel perspectives**. *Annu Rev Psychol* 2005, **56**:365–392. https://doi.org/10.1146/annurev.psych.56.091103.070141

2. Curry OS, Rowland LA, Van Lissa CJ, Zlotowitz S, Mcalaney J, Whitehouse H: **Happy to help? A systematic review and meta-analysis of the effects of performing acts of kindness on the well-being of the actor**. *J Exp Soc Psychol* 2018, **76**:320–329. https://doi.org/10.1016/j.jesp.2018.02.014

3. Hui BPH, Ng JCK, Berzaghi E, Cunningham-Amos LA, Kogan A: **Rewards of kindness? A meta-analysis of the link between prosociality and well-being**. *Psychol Bull* 2020, **146**:1084–1116. https://doi.org/10.1037/bul0000298

4. Yaakobi E, Weisberg J: **Organizational citizenship behavior predicts quality, creativity, and efficiency performance: The roles of occupational and collective efficacies**. *Front Psychol* 2020, **11**:758. https://doi.org/10.3389/fpsyg.2020.00758

5. Góis AR, Santos FP, Pacheco JM, Santos FC: **Reward and punishment in climate change dilemmas**. *Sci Rep* 2019, **9**:16193. https://doi.org/10.1038/s41598-019-52524-8

** A mathematical model on a N-player Collective-Risk dilemma (CRD) reveals that rewards are essential to initiate cooperation when cooperation level is low, whereas punishment helps maintain cooperation when a certain level of cooperation is reached. Thus, the best strategy to solve challenging social dilemmas is to use reward and punishment in tandem.

6. Jin S, Balliet D, Romano A, Spadaro G, van Lissa CJ, Agostini M, Bélanger JJ, Gützkow B, Kreienkamp J, Leander NP, et al.: **Intergenerational conflicts of interest and prosocial behavior during the COVID-19 pandemic**. *Pers Individ Dif* 2021, **171**:110535. https://doi.org/10.1016/j.paid.2020.110535

7. Balliet D, Mulder LB, Van Lange PAM: **Reward, punishment, and cooperation: A meta-analysis**. *Psychol Bull* 2011, **137**:594–615. https://doi.org/10.1037/a0023489

8. van Dijk E, Molenmaker WE, de Kwaadsteniet EW: **Promoting cooperation in social dilemmas: The use of sanctions**. *Curr Opin Psychol* 2015, **6**:118–122. https://doi.org/10.1016/j.copsyc.2015.07.006

9. Ohtsuki H, Iwasa Y, Nowak MA: **Indirect reciprocity provides only a narrow margin of efficiency for costly punishment**. *Nature* 2009, **457**:79–82. https://doi.org/10.1038/nature07601

10. Raihani NJ, Bshary R: **Punishment: One tool, many uses**. *Evol Hum Sci* 2019, **1**:e12. https://doi.org/10.1017/ehs.2019.12

11. Fehr E, Gächter S.: **Altruistic punishment in humans**. *Nature* 2002, **415**:137–140. https://doi.org/10.1038/415137a

12. Ozono H, Kamijo Y, Shimizu K: **The role of peer reward and punishment for public goods problems in a localized society**. *Sci Rep* 2020, **10**:8211. https://doi.org/10.1038/s41598-020-64930-4

13. Noussair CN, van Soest D, Stoop J: **Punishment, reward, and cooperation in a**

framed field experiment. *Soc Choice Welfare* 2015, **45**:537–559. https://doi.org/10.1007/s00355-014-0841-8

14. Dreber A, Rand DG, Fudenberg D, Nowak MA: **Winners don't punish**. *Nature* 2008, **452**:348–351. https://doi.org/10.1038/nature06723

15. Muñoz-Herrera M, Nikiforakis N: *Experimental evidence shows that negative motive attribution drives counter- punishment*. 2020. https://nyuad.nyu.edu/content/dam/nyuad/academics/divisions/social-science/working-papers/2020/0056.pdf

16. Gross J, Méder ZZ, Okamoto-Barth S, Riedl A: **Building the Leviathan-Voluntary centralisation of punishment power sustains cooperation in humans**. *Sci Rep* 2016, **6**:20767. https://doi.org/10.1038/srep20767

17. Heine F, Strobel M: **Reward and punishment in a team contest**. *PLoS One* 2020, **15**:e0236544. https://doi.org/10.1371/journal.pone.0236544

18. Chao M: **Demotivating incentives and motivation crowding out in charitable giving**. *Proc Natl Acad Sci U S A* 2017, **114**:7301–7306. https://doi.org/10.1073/pnas.1616921114

19. Gneezy U, Meier S, Rey-Biel P: **When and why incentives (don't) work to modify behavior**. *J Econ Perspect* 2011, **25**:191–210. https://doi.org/10.1257/jep.25.4.191

20. Okada I: **A review of theoretical studies on indirect reciprocity**. *Games* 2020, **11**:27. https://doi.org/10.3390/g11030027

21. Panchanathan K, Boyd R: **A tale of two defectors: The importance of standing for evolution of indirect reciprocity**. *J Theor Biol* 2003, **224**:115–126.

https://doi.org/10.1016/S0022-5193(03)00154-1

22. Yamamoto H, Suzuki T, Umetani R: **Justified defection is neither justified nor unjustified in indirect reciprocity**. *PLoS One* 2020, **15**:e0235137. https://doi.org/10.1371/journal.pone.0235137

23. Chen X, Sasaki T, Brännström Å, Dieckmann U: **First carrot, then stick: How the adaptive hybridization of incentives promotes cooperation**. *J R Soc Interface* 2015, **12**:20140935. https://doi.org/10.1098/rsif.2014.0935

24. Heffner J, FeldmanHall O: **Why we don’t always punish: Preferences for non-punitive responses to moral violations**. *Sci Rep* 2019, **9**:13219. https://doi.org/10.1038/s41598-019-49680-2

25. Jordan JJ, Hoffman M, Bloom P, Rand DG: **Third-party punishment as a costly signal of trustworthiness**. *Nature* 2016, **530**:473–476. https://doi.org/10.1038/nature16981

26. Molenmaker WE, de Kwaadsteniet EW, van Dijk E: **The impact of personal responsibility on the (un)willingness to punish non-cooperation and reward cooperation**. *Organ Behav Hum Decis Process* 2016, **134**:1–15. https://doi.org/10.1016/j.obhdp.2016.02.004

27. de Kwaadsteniet EW, Rijkhoff SAM, Dijk E Van: **Equality as a benchmark for third-party punishment and reward: The moderating role of uncertainty in social dilemmas**. *Organ Behav Hum Decis Process* 2013, **120**:251–259. https://doi.org/10.1016/j.obhdp.2012.06.007

28. Inesi ME, Adams GS, Gupta A: **When it pays to be kind: The allocation of indirect**

reciprocity within power hierarchies. *Organ Behav Hum Decis Process* 2021, **165**:115–126. https://doi.org/10.1016/j.obhdp.2021.04.005

29. Dhaliwal NA, Patil I, Cushman F: **Reputational and cooperative benefits of third-party compensation**. *Organ Behav Hum Decis Process* 2021, **164**:27–51. https://doi.org/10.1016/j.obhdp.2021.01.003

** This research examined the consequences of punishing the perpetrators or compensating the victims across 24 studies involving various norm violation contexts. Overall, compensating victims leads to more reputational and partner choice benefits than punishing perpetrators, and individuals' decision to compensate or punish depends on injunctive and descriptive norms.

30. Lee Y, Warneken F: **Children's evaluations of third-party responses to unfairness: Children prefer helping over punishment**. *Cognition* 2020, **205**:104374. https://doi.org/10.1016/j.cognition.2020.104374

31. Ozono H, Watabe M: **Reputational benefit of punishment: Comparison among the punisher, rewarder, and non-sanctioner**. *Lett Evol Behav Sci* 2012, **3**:21–24. https://doi.org/10.5178/lebs.2012.22

32. de Quervain DJF, Fischbacher U, Treyer V, Schellhammer M, Schnyder U, Buck A, Fehr E: **The neural basis of altruistic punishment**. *Science* 2004, **305**:1254–1258. https://doi.org/10.1126/science.1100735

33. Lopez LD, Moorman K, Schneider S, Baker MN, Holbrook C: **Morality is relative: Anger, disgust, and aggression as contingent responses to sibling versus acquaintance harm**. *Emotion* 2021, **21**:376–390. https://doi.org/10.1037/emo0000707

34. Molho C, Tybur JM, Van Lange PAM, Balliet D: **Direct and indirect punishment of norm violations in daily life**. *Nat Commun* 2020, **11**:3432. https://doi.org/10.1038/s41467-020-17286-2

** This longitudinal study on interventions to norm violations in daily life reveals that whether people tend to directly confront the offender or gossip about the offender depends on situational cues (e.g., power asymmetry, moral wrongness of norm violations, and valuation of offenders) about the benefits of changing others' behavior and the risk of retaliation.

35. Tybur JM, Molho C, Cakmak B, Cruz TD, Singh GD, Zwicker M: **Disgust, anger, and aggression: Further tests of the equivalence of moral emotions**. *Collabra Psychol* 2020, **6**:34. https://doi.org/10.1525/collabra.349

36. Ginther MR, Hartsough LES, Marois R: **Moral outrage drives the interaction of harm and culpable intent in third-party punishment decisions.** *Emotion* 2021, https://doi.org/10.1037/emo0000950

37. Wang CS, Sivanathan N, Narayanan J, Ganegoda DB, Bauer M, Bodenhausen G V., Murnighan K: **Retribution and emotional regulation: The effects of time delay in angry economic interactions**. *Organ Behav Hum Decis Process* 2011, **116**:46–54. https://doi.org/10.1016/j.obhdp.2011.05.007

38. Molenmaker WE, de Kwaadsteniet EW, van Dijk E: **The effect of decision timing on the willingness to costly reward cooperation and punish noncooperation: Sanctioning the past, the present, or the future**. *J Behav Decis Mak* 2019, **32**:241–254. https://doi.org/10.1002/bdm.2110

39. Pfattheicher S, Sassenrath C, Keller J: **Compassion magnifies third-party**

punishment. *J Pers Soc Psychol* 2019, **117**:124–141. https://doi.org/10.1037/pspi0000165

40. Vayness J, Duong F, DeSteno D: **Gratitude increases third-party punishment**. *Cogn Emot* 2020, **34**:1020–1027. https://doi.org/10.1080/02699931.2019.1700100

41. Delton AW, Krasnow MM: **The psychology of deterrence explains why group membership matters for third-party punishment**. *Evol Hum Behav* 2017, **38**:734–743. https://doi.org/10.1016/j.evolhumbehav.2017.07.003

42. Krasnow MM, Delton AW, Cosmides L, Tooby J: **Looking under the hood of third-party punishment reveals design for personal benefit**. *Psychol Sci* 2016, **27**:405–418. https://doi.org/10.1177/0956797615624469

43. Pedersen EJ, McAuliffe WHB, Shah Y, Tanaka H, Ohtsubo Y, McCullough ME: **When and why do third parties punish outside of the lab? A cross-cultural recall study**. *Soc Psychol Personal Sci* 2020, **11**:846–853. https://doi.org/10.1177/1948550619884565

** This cross-cultural recall study reveals that in real-life social interactions, third-party observers’ anger and punishment behavior toward the transgressors depend on how much they value the welfare of the transgressor and the victim.

44. Gordon DS, Lea SEG: **Who punishes ? The status of the punishers affects the perceived success of, and indirect benefits from, “moralistic” punishment**. *Evol Psychol* 2016, **14**:1–14. https://doi.org/10.1177/1474704916658042

45. Redhead D, Dhaliwal N, Cheng JT: **Taking charge and stepping in: Individuals who punish are rewarded with prestige and dominance**. *Soc Personal Psychol*

Compass 2021, **15**:e12581. https://doi.org/10.1111/spc3.12581

46. Jordan JJ, McAuliffe K, Warneken F: **Development of in-group favoritism in children's third-party punishment of selfishness**. *Proc Natl Acad Sci U S A* 2014, **111**:12710–12715. https://doi.org/10.1073/pnas.1402280111

47. Gneezy A, Fessler DMT: **Conflict, sticks and carrots: War increases prosocial punishments and rewards**. *Proc R Soc B Biol Sci* 2012, **279**:219–223. https://doi.org/10.1098/rspb.2011.0805

48. Mifune N, Li Y, Okuda N: **The evaluation of second- and third-party punishers**. *Lett Evol Behav Sci* 2020, **11**:6–9. https://doi.org/10.5178/lebs.2020.72

49. Deutchman P, Bračič M, Raihani N, McAuliffe K: **Punishment is strongly motivated by revenge and weakly motivated by inequity aversion**. *Evol Hum Behav* 2021, **42**:12–20. https://doi.org/10.1016/j.evolhumbehav.2020.06.001

50. Van Lange PAM, Ouwerkerk JW, Tazelaar MJA: **How to overcome the detrimental effects of noise in social interaction: The benefits of generosity**. *J Pers Soc Psychol* 2002, **82**:768–780. https://doi.org/10.1037/0022-3514.82.5.768

51. de Kwaadsteniet EW, Kiyonari T, Molenmaker WE, van Dijk E: **Do people prefer leaders who enforce norms? Reputational effects of reward and punishment decisions in noisy social dilemmas**. *J Exp Soc Psychol* 2019, **84**:103800. https://doi.org/10.1016/j.jesp.2019.03.011

** Findings from three studies reveal that leaders' decisions to reward and punish positively affect their reputations in noise-free situations, but their decisions to reward more positively affect their reputations than decisions to punish when there is noise in the situation.

52. Grechenig KR, Nicklisch A, Thöni C: **Punishment despite reasonable doubt - A public goods experiment with uncertainty over contributions**. *J Empir Leg Stud* 2010, **7**:847–867. https://doi.org/10.1111/j.1740-1461.2010.01197.x

53. Dong Y, Sasaki T, Zhang B: **The competitive advantage of institutional reward**. *Proc R Soc B Biol Sci* 2019, **286**:20190001. https://doi.org/10.1098/rspb.2019.0001

54. Hess NH, Hagen EH: **Psychological adaptations for assessing gossip veracity**. *Hum Nat* 2006, **17**:337–354. https://doi.org/10.1007/s12110-006-1013-z

55. Ambrus A, Greiner B: **Individual, dictator, and democratic punishment in public good games with perfect and imperfect observability**. *J Public Econ* 2019, **178**:104053. https://doi.org/10.1016/j.jpubeco.2019.104053

56. Chen H, Zeng Z, Ma J: **The source of punishment matters: Third-party punishment restrains observers from selfish behaviors better than does second-party punishment by shaping norm perceptions**. *PLoS One* 2020, **15**:e0229510. https://doi.org/10.1371/journal.pone.0229510

* Two experiments reveal that third-party punishment is more effective than second-party punishment in enhancing other potential partners' normative beliefs (i.e., beliefs that unfair distributive behavior is unusual and unacceptable) and prosocial behavior.

57. Pfattheicher S, Böhm R, Kesberg R: **The advantage of democratic peer punishment in sustaining cooperation within groups**. *J Behav Decis Mak* 2018, **31**:562–571. https://doi.org/10.1002/bdm.2050

58. Balafoutas L, Nikiforakis N, Rockenbach B: **Direct and indirect punishment among strangers in the field**. *Proc Natl Acad Sci U S A* 2014, **111**:15924–15927.

https://doi.org/10.1073/pnas.1413170111

59. Grimalda G, Pondorfer A, Tracer DP: **Social image concerns promote cooperation more than altruistic punishment**. *Nat Commun* 2016, **7**:12288. https://doi.org/10.1038/ncomms12288

60. Wu J, Balliet D, Van Lange PAM: **Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation**. *Sci Rep* 2016, **6**:23919. https://doi.org/10.1038/srep23919

61. Fang Y, Benko TP, Perc M, Xu H, Tan Q: **Synergistic third-party rewarding and punishment in the public goods game**. *Proc R Soc A Math Phys Eng Sci* 2019, **475**:20190349. https://doi.org/10.1098/rspa.2019.0349

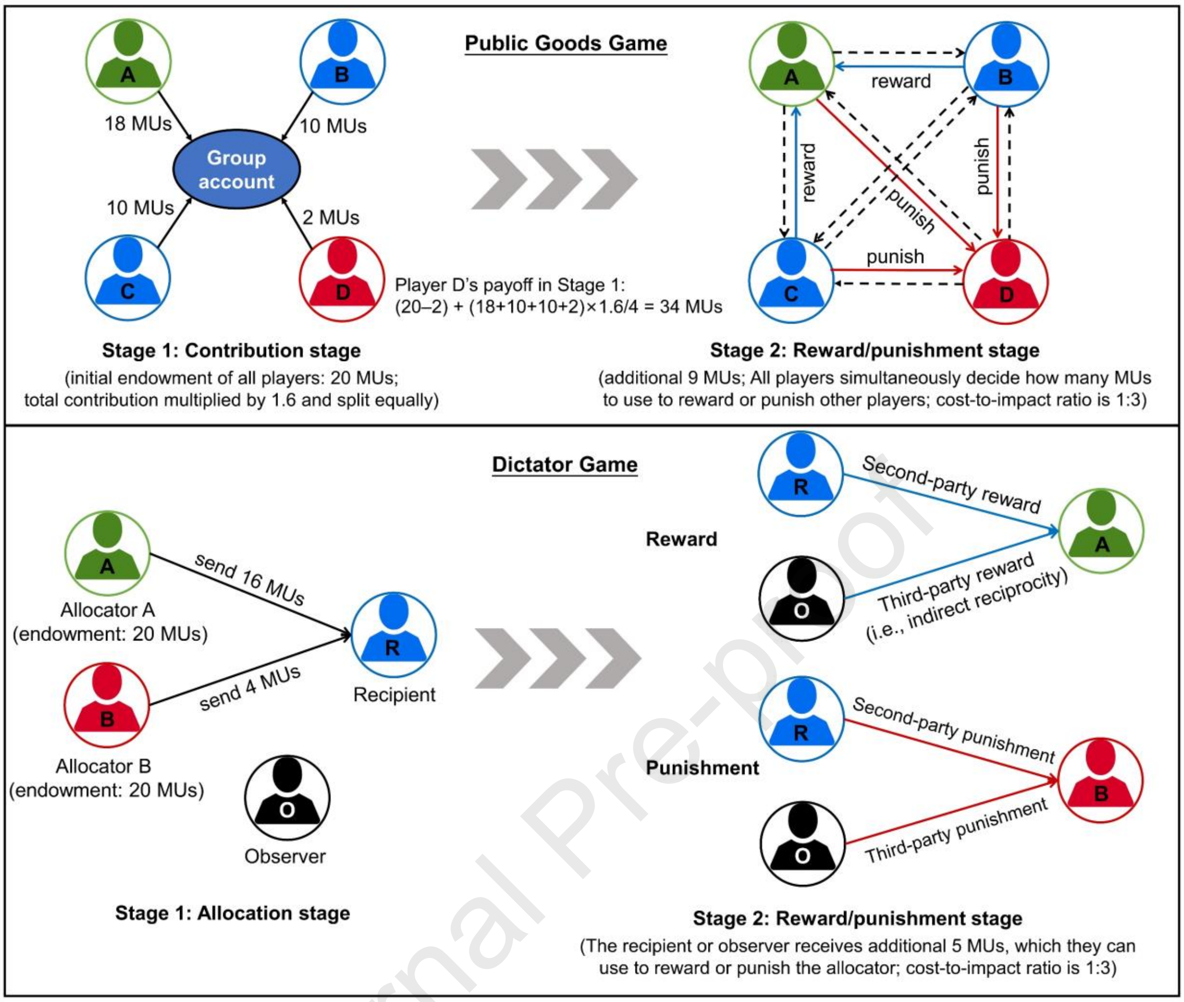
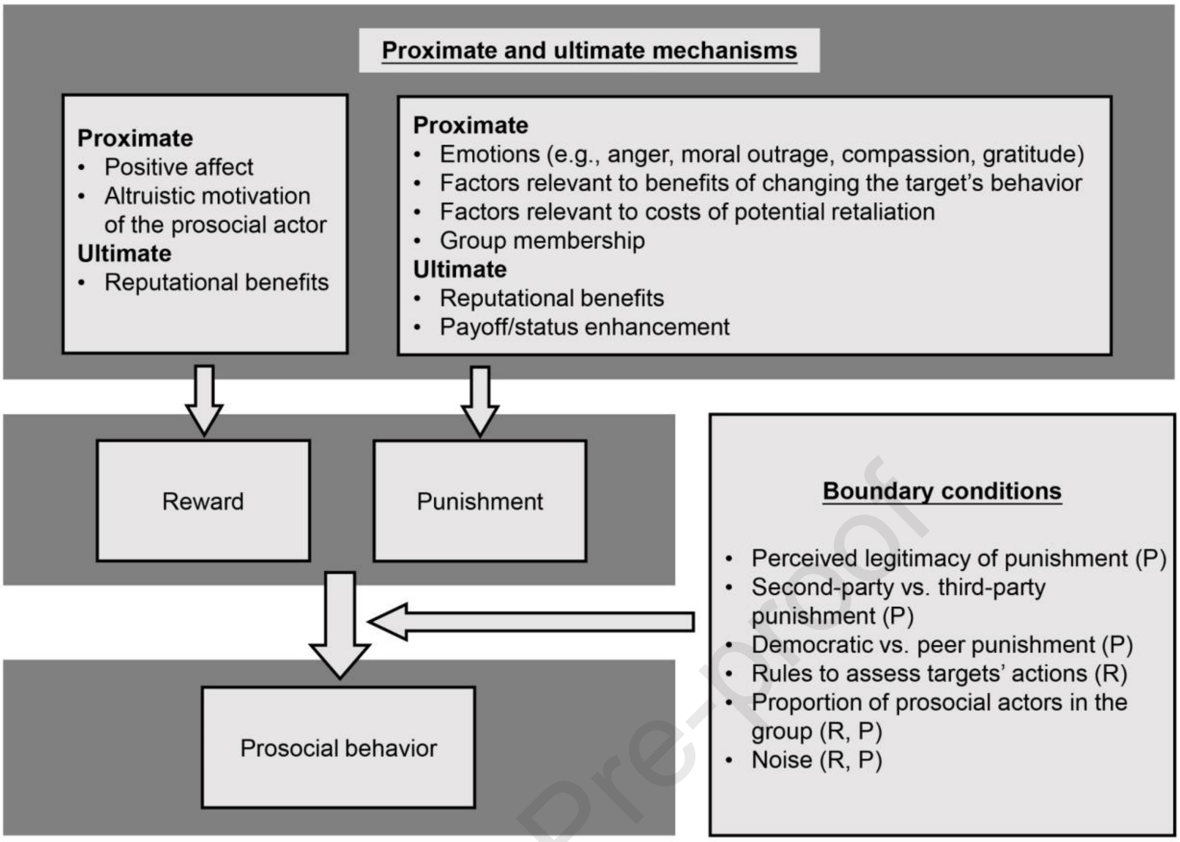

Public Goods Game
A
B
C
D
18 MUs
10 MUs
10 MUs
2 MUs
Group account
Player D's payoff in Stage 1:
(20–2) + (18+10+10+2)×1.6/4 = 34 MUs
Stage 1: Contribution stage
(initial endowment of all players: 20 MUs;
total contribution multiplied by 1.6 and split equally)
reward
reward
punish
punish
punish
Stage 2: Reward/punishment stage
(additional 9 MUs; All players simultaneously decide how many MUs
to use to reward or punish other players; cost-to-impact ratio is 1:3)
Dictator Game
Allocator A
(endowment: 20 MUs)
Allocator B
(endowment: 20 MUs)
send 16 MUs
send 4 MUs
R
Recipient
O
Observer
Stage 1: Allocation stage
Reward
Second-party reward
Third-party reward
(i.e., indirect reciprocity)
Punishment
Second-party punishment
Third-party punishment
Stage 2: Reward/punishment stage
(The recipient or observer receives additional 5 MUs, which they can
use to reward or punish the allocator; cost-to-impact ratio is 1:3)

Figure 1. Illustrations of payoff structures in a public goods game and a dictator game, and the administration of reward and punishment in these games. Reward: assigning 1 MU to a target costs the rewarder 1 MU, and benefits the target by 3 MUs; Punishment: assigning 1 MU to a target costs the punisher 1 MU, and costs the target by 3 MUs. MU = monetary unit.



Proximate and ultimate mechanisms
Proximate
• Positive affect
• Altruistic motivation of the prosocial actor
Ultimate
• Reputational benefits
Proximate
• Emotions (e.g., anger, moral outrage, compassion, gratitude)
• Factors relevant to benefits of changing the target's behavior
• Factors relevant to costs of potential retaliation
• Group membership
Ultimate
• Reputational benefits
• Payoff/status enhancement
Reward
Punishment
Prosocial behavior
Boundary conditions
• Perceived legitimacy of punishment (P)
• Second-party vs. third-party punishment (P)
• Democratic vs. peer punishment (P)
• Rules to assess targets' actions (R)
• Proportion of prosocial actors in the group (R, P)
• Noise (R, P)

Figure 2. Overview of the proximate and ultimate mechanisms of rewarding and punishing decisions, as well as the boundary conditions for the effects of reward and punishment on prosocial behavior. R = applies to reward, P = applies to punishment.

Credit Author Statement

Junhui Wu: Conceptualization, Writing – original draft, Visualization. **Shenghua Luan:** Writing – review & editing. **Nichola Raihani:** Conceptualization, Writing – review & editing.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: