The reputation of punishers

Nichola J Raihani[1] and Redouan Bshary[2]

1. Department of Genetics, Evolution and Environment, University College London, WC1E 6BT, UK.
2. Institut de Biologie, Eco-Ethologie, Université de Neuchâtel, Neuchâtel, CH-2000, Switzerland.

Corresponding author: Raihani, NJ (nicholaraihani@gmail.com)
Keywords: punishment, reputation, fairness, partner choice

Punishment is a potential mechanism to stabilise cooperation between self-regarding agents. Theoretical and empirical studies on the importance of a punitive reputation have yielded conflicting results. Here, we propose that a variety of factors interact to explain why a punitive reputation is sometimes beneficial and sometimes harmful. We predict that benefits are most likely to occur in forced play scenarios and in situations where punishment is the only means to convey an individual's cooperative intent and willingness to uphold fairness norms. In contrast, if partner choice is possible and an individual's cooperative intent can be inferred directly, then individuals with a non-punishing cooperative reputation should typically be preferred over punishing cooperators.

The puzzle of punishment

Punishment is a mechanism that can promote cooperation where individuals would otherwise be tempted to cheat [1-3]. Humans are apparently willing to engage in costly punishment of others and derive subjective pleasure from doing so [4, 5]. However, since punishment imposes immediate costs on punishers, two important issues arise that are central in both theoretical and empirical studies. First, one has to elucidate how punishers are eventually compensated for their investment. In repeated games, punishers can directly benefit if the target behaves more cooperatively in the future as a consequence of being punished. This logic applies to both 2-player and n-player games [2, 6], though in the latter punishment might be under negative frequency dependence as exemplified in a volunteer's dilemma [7-9]. In one-shot games, it has been argued that punishment can only evolve through higher level population processes (e.g. [10, 11]) that are ultimately based on increasing the punisher's inclusive fitness [12]. The second issue concerns the

fact that punishment by definition destroys value [13], even more so if it leads to counter-punishment rather than cooperation, as frequently observed in experiments on humans (e.g. [14-18]). While the theoretical literature provides conflicting results on the evolvability of vendetta strategies as opposed to co-evolution between cooperating and punishing defectors [19-21], one has to ask under what conditions punishment could be favoured over alternative non-destructive or even value-building control mechanisms like reciprocal defection, rewarding co-operators or compensating victims (e.g. [13, 16, 22-25]). Here, we are particularly concerned with peer punishment and ignore the evolution of centralized punishment institutions, although we note that similar arguments to those that we make about peer punishment might also apply to centralized punishment (e.g. see [1, 26]). We explore under what conditions a punitive reputation could facilitate or hinder the evolution and maintenance of punishment as a partner control mechanism. As all empirical evidence on the reputation consequences of punishment is limited to humans, we do not discuss punishment in other species.

The reputation of punishers

It has been argued that many of the difficulties in reconciling the immediate costs of punishing with ultimate fitness benefits to the punisher can be overcome if interactions are not anonymous as punishers can then benefit from acquiring a punitive reputation. Reputation generally offers one major solution to the question of why humans cooperate: helping others improves an individual's reputation, which in turn increases the probability of receiving help from observers or being chosen as cooperation partners [27-29]. In much the same way, theoretical models of 2-player and n-player games have shown that when individuals are forced to interact with one another and can infer how the partner is likely to behave by observing whether the partner punished in previous interaction(s), then punishment can evolve via direct reputation-based benefits to the punisher (e.g. [11, 21, 22, 24, 30, 31]). According to these models, strategies that respond conditionally to the punitive reputation of the partner (by cooperating when paired with a punisher but defecting when paired with a non-punisher) typically outperform unconditionally cooperative or defecting strategies because individuals can avoid the risk of being punished but accrue the benefits of defecting otherwise. Punishers benefit from investing in responsible punishment because being identified as a punisher reduces the risk that an opportunistic partner will defect. The possibility for individuals to benefit from acquiring a punitive reputation catalyses the emergence of responsible punishment while preventing the spread of antisocial and spitefully punishing strategies (see [21]; Glossary). The latter

would otherwise be expected to outperform responsible punishment strategies (e.g. [19, 20]).

Empirical data on the reputation consequences of punishment are quite mixed. While some studies support the idea proposed by the theoretical models that individuals benefit from a punitive reputation because future partners are deterred from cheating (e.g., [32, 33]), findings from other studies are less straightforward to interpret. For example, punishers have been demonstrated to both advertise [34, 35] and hide [36] punitive behaviour, the former findings hinting that punishment yields reputation benefits while the latter indicates otherwise. While punishers are generally trusted more than non-punishers (e.g. [37, 38]) various studies have failed to provide evidence that punishers are liked or rewarded for their investment [32, 37, 39-42]. Nevertheless, a recent article based on evaluations of vignettes indicated that third-party punishers were judged as more likeable than those who either did not apprehend cheats or who were the victim of the cheat themselves [43]; while another empirical study has also shown that bystanders are more likely to reward third-party punishers than individuals who take no action in response to a cheat [Raihani & Bshary, in review]. A possible conclusion to draw from the findings of these models and empirical studies is that individuals can benefit from a punitive reputation in some contexts but not others. Here, we provide a conceptual framework (Figure 1; Box 1) that integrates existing theoretical and empirical knowledge when possible but also makes predictions for hitherto unexplored scenarios (Table 1). While necessarily somewhat speculative in nature, our framework will hopefully inspire both further experimentation and modelling. We suggest that to understand the reputation consequences of punishment it is necessary to first determine what kind of information can be extracted by observers regarding the potential motive underpinning the decision to punish before assessing when, and why, such a reputation is beneficial or harmful to the punisher. As argued by [44], motives can be spiteful, self-serving, other-regarding or a mixture, and some situations might allow less ambiguous assessments than others. Therefore, punishment can be more or less indicative of an individual's cooperative tendencies and fairness preferences (as suggested by [38, 44]). As a consequence, a punitive reputation might have diverse effects on the future behaviour of observers, both regarding their willingness to engage in interactions with the punisher if partner choice is an option, and their willingness to cooperate in forced play scenarios. A central unresolved question is whether a punitive reputation would ever induce observers to voluntarily help or reward punishers in the absence of any punishment threat.

What does punishment signal?

Punishment might always signal that the punisher is willing to incur costs to harm others. However, on its own this information is not helpful to observers in predicting how punishers might behave in future. Instead, we suggest that observers need to know about the context in which punishment occurred in order to infer what motivated the punishment decision and how the punisher is therefore likely to behave in subsequent interactions. We predict that punishment can affect reputation in two ways: it can signal that the punisher is competitive or, under more restricted circumstances, can signal the punisher's cooperative intent. We predict that observers only infer cooperative intent from punishment if competitive motives can be excluded. To verify this prediction and thereby understand the reputation consequences of punishment, it is crucial for both empirical studies and theoretical models to vary the nature of the game in which a punitive reputation is built.

If a punisher is directly and actively involved in the game where punishment occurs, observers can gain information about the punisher's own behaviour (cooperate or defect), whether the punishment is potentially self-serving (one-off interaction versus iterated game), and whether the punisher is the only beneficiary of their investment or whether others benefit as well (2 player or n-player game). If an individual is directly but passively involved in the game (e.g. a potential recipient of helping, as in [32]), observers can only assess whether a punisher acted against a helping or a non-helping partner and whether there are potential personal or shared future benefits of doing so, but not whether the punisher would cooperate or defect herself. Finally, if an individual was a third-party, observers can assess whether the punishment was justified or not but cannot assess whether the individual would cooperate or defect herself. We predict that punishment is most likely to serve as a proxy for the cooperative tendency of the punisher when (i) punishment is aimed at defectors and not cooperators; (ii) the cooperative tendency of the punisher cannot be directly observed; and (iii) the punisher was not the primary victim of the cheat (Figure 1; Box 1). Some empirical evidence supports these predictions: antisocial punishment and retaliatory punishment (where the punisher is the victim of the cheat) have both been shown to correlate with competitive, rather than cooperative, motives [45-48]. Data are currently lacking as to whether investment in third-party punishment is a reliable signal of cooperative tendency, as predicted (but see [Jordan et al., under review] for supportive evidence).

We propose that another important potential signalling aspect of punishment is linked to its efficiency. Empirical studies typically use a fee to fine ratio that is larger than one (e.g. [4, 14, 16 - 18, 25, 32, 34 - 41]). This approach has been adopted in some recent models [30, 31] while early theoretical papers show that it is not a necessity for the evolution of punitive reputations [11, 22]. A fee to fine ratio larger than one indicates that punishing is cheaper than being punished and, for ease, we henceforth refer to this as an 'efficient' punishment regime. This assumption certainly helps the evolution of punishment when individuals are forced to interact with one another and the reputation consequences of punishment induce opportunistic cooperators (individuals who cooperate only under the threat of punishment) to cooperate. However, since efficient punishment always improves the punisher's payoff's relative to those of the target, we suggest that punishers always face the observers' doubt concerning the moral value of the act [49, 50]. Even for third-party punishment, where the punisher was not the victim of the defection, observers could conclude that the motivation behind the punishment is jealousy and the punisher's aim is to reduce the asymmetry between their own and the defector's payoff (e.g. [51]). Previous work has shown that third-party punishers that incur high costs (relative to their initial endowment) to punish cheats are preferred over those that incur lower costs [38], which is consistent with the idea that observers take the relative cost of punishment into account when evaluating the punisher. When punishment is efficient, we predict that third-party punishers will be evaluated less positively than third-party individuals who help the victim instead, a prediction which is supported by a recent experiment using a 4:1 fee-to-fine ratio for both third-party punishment and third-party helping [Raihani & Bshary, under review]. Fittingly, recent studies have also shown that that when third-parties are given the option to either punish a cheat or to compensate the victim, more people choose the latter [25, 52; Raihani & Bshary, under review], with the option to reverse the fortunes of the cheat and victim, when available, being the most popular [53].

How does the punitive signal affect the behaviour of observers?

We have argued that punishment could serve as a competitive signal (highlighting the punisher's willingness to incur costs to admonish others) and, in more restricted circumstances, as a signal of cooperative intent. These two different signals are likely to have different impacts on the emotions induced in observers and - accordingly - will determine whether the punitive reputation is helpful or harmful to the punisher. To determine the reputation consequences of punishment the next step is to investigate the rules of the second game in which former observers interact with initial players. We

consider three factors as potentially important. First, what kind of game is played? Second, can observers choose their interaction partners? Third, is punishment an option in the second game? The existing theoretical and empirical literature agrees on one main result. Where punishment is justified, efficient punishment can stabilise cooperation and typically outcompetes other partner control mechanisms if individuals are forced to interact with one another and punishment will be possible in the next game [9, 17, 18, 20, 26-28]. Crucially, however, although most experiments reveal that efficient punishers tend to be trusted by others (e.g. [37, 38]), they are not necessarily liked more than cooperative non-punishers [37, 39] or rewarded for their investment [32, 39 - 42]. Taken together, these results lead us to conclude that it is fear (but not love) induced in observers that renders a punitive reputation beneficial to the punisher in forced-play games with punishment options. The question of whether a punitive reputation can induce love in observers, leading to voluntary help or cooperation even when punishment is no longer a threat, is still open. Recent evidence suggests that third-party punishers are liked more than individuals that do not punish cheats [43] but it is not known whether uninvolved bystanders would pay to reward these individuals for their punitive investment (but see [Raihani & Bshary, under review]).

To our knowledge, no theoretical study and few empirical studies have asked how a punisher's reputation affects the possibility to be chosen as a partner. Assuming that individuals prefer to choose cooperative partners for interactions, we expect that punishers will only be preferred over non-punishers where punishment reliably signals cooperative intent *in the absence* of more direct information on a potential partner's cooperative behaviour. In the absence of opportunities to perform positive actions, third-party punishers are preferred over non-punishers for interactions (e.g. [38]) and are more likely to punish when their decision is made known to others (e.g. [34, 35]). Conversely, punishers might be less preferred as partners when cooperative intent can be signalled through positive actions instead, or when punishment can be interpreted as competitive. The findings of a recent study by Rockenbach & Milinski [36] partially support this prediction. In this study, individuals that signalled cooperative intent through contributions to the public good paid to conceal (severe) punishment from prospective observing partners, hinting that people believed that their punitive decisions would be evaluated negatively. Observers joined the group in the next round and could exclude one player from the initial team of four from the interaction. Apparently, punitive behaviour by players did not affect the observers' choice of who to exclude. Instead, observers excluded

partners based on low contributions to the public good rather than punishment [36]. Thus, while this study shows that cooperative individuals are preferred over non-cooperative individuals it does not tell us how cooperative punishers are evaluated relative to cooperative non-punishers. To properly evaluate the reputation consequences of punishment, while holding information about the partner's cooperativeness constant, would require a new study where the observer is forced to actively choose one player out of four rather than excluding one player out of four. We predict that when punishment is efficient, observers will prefer non-punishing cooperators if a) they themselves are defectors; b) the game to be played is a 2-player game; or c) the observer knows that all other partners are cooperative in an n-player game. The latter two predictions are based on standard arguments in the theoretical literature that both execution and perception errors are part of life [54, 55], which in turn would lower the observer's payoff if he chooses a partner who has punishment in his repertoire. Conversely, we predict that observers will prefer to include punishing cooperators in n-player interactions when there is a risk that other members of the future group are defectors or opportunistic co-operators. The optimal number of punishers to include is likely to depend on whether linear or non-linear benefits of punishment are assumed [7]. In the absence of punishment options, we assume that models would predict a positive effect of a punitive reputation only in the absence of direct information on cooperative tendencies, and that this effect will be more pronounced if punishment is inefficient since inefficient punishment more easily allows competitive motives to be ruled out. When both punitive and cooperative information is present - but punishment is no longer an option - punitive scores should be ignored. However, as our reputation concept is psychological in nature, we would also predict that in empirical studies a purely cooperative reputation outcompetes a punitive reputation that is based on fear, while justified and inefficient punishment might cause no disadvantage or potentially even an advantage in terms of partner choice. It is clear that more theoretical and empirical work is needed to clarify the circumstances under which a punitive reputation is beneficial, particularly when partner choice is possible (Table 1).

Concluding remarks

Our framework identifies key parameters that might affect the reputation effects of punishment. We expect that observers will infer competitive motives of punishers unless these can be ruled out. According to our predictions, competitive punishers might be respected but also feared by observers, whereas cooperative punishers are more likely to be loved. Being feared should be beneficial in forced play games but could be harmful if

observers can avoid the punisher for future interactions. Conversely, punishers that are truly loved could be evaluated in a similar manner to helpful individuals and thus preferentially selected for interactions by observers. Partner choice might therefore hold the key for understanding why individuals sometimes advertise punishment but conceal their actions at other times and the field would benefit hugely from more theoretical models and empirical studies exploring this possibility. Perhaps most importantly, whereas punishers might gain benefits from being feared in enforced games, it is still an open question whether there are conditions where they will benefit from being loved.

References

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

2. Clutton-Brock, T. H., and Parker GA. (1995) Punishment in animal societies. Nature 373, 209–216

3. Raihani, N.J., Thornton, A., and Bshary, R. (2012) Punishment and cooperation in nature. Trends Ecol. Evol. 27, 288–295

4. Fehr, E. and Gachter, S. (2002) Altruistic punishment in humans. Nature 415, 137–140

5. De Quervain, D., et al. (2004) The neural basis of altruistic punishment. Science 305, 1254-1258

6. Bshary, A., and Bshary, R. (2010) Self-serving punishment of a common enemy creates a public good in reef fishes. Curr. Biol. 20, 2032–2035

7. Raihani, N. J., and Bshary, R. (2011) The evolution of punishment in n-player public goods games: a volunteer's dilemma. Evolution 65, 2725-2728

8. Archetti, M. (2009) Cooperation as a volunteer’s dilemma and the strategy of conflict in public goods games. J. Evol. Biol. 22, 2192–2200

9. Diekmann, A. (1985) Volunteer's dilemma. J. Confl. Resolut. 29, 605-610

10. Boyd, R. et al. (2003) The evolution of altruistic punishment. Proc. Natl. Acad. Sci. U.S.A. 100, 3531-3535

11. Brandt, H., Hauert, C., and Sigmund, K. (2003) Punishment and reputation in spatial public goods games. Proc. R. Soc. Lond. B. 270, 1099–1104

12. Lehmann, L. et al. (2007) Strong reciprocity or strong ferocity? A population genetic view of the evolution of altruistic punishment. Am. Nat. 170, 21–36

13. Ohtsuki, H., Iwasa, Y., and Nowak, M. (2009) Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature, 457, 79–82

14. Herrmann, B., Thöni, C. & Gächter, S. (2008) Antisocial punishment across societies.

Science 319, 1362-1367
15. Janssen, M., and Bushman, C. (2008) Evolution of cooperation and altruistic punishment when retaliation is possible. J. Theor. Biol. 254, 541–545
16. Dreber, A. et al. (2008) Winners don't punish. Nature 452, 348–351
17. Nikiforakis, N. & Engelmann, J. (2011) Altruistic punishment and the threat of feuds. J Econ. Behav. Organ. 78, 319-332
18. Fehl, K., Sommerfeld, R.D., Semmann, D., Krambeck, H.J. & Milinski, M. (2012) I dare you to punish me : vendettas in games of cooperation. PLoS ONE, 7, e4503
19. Rand, D. G. et al. (2010) Anti-social punishment can prevent the co-evolution of punishment and cooperation. J. Theor. Biol. 265, 624–632
20. Rand, D. G., and Nowak, M. A. (2011) The evolution of antisocial punishment in optional public goods games. Nat. Comm. ncomm1442
21. Hilbe, C., and Traulsen, A. (2012) Emergence of responsible sanctions without second-order free riders, antisocial punishment or spite. Sci. Rep. 458, srep00458
22. Sigmund, K., Hauert, C., and Nowak, M. A. (2001) Reward and punishment. Proc. Natl. Acad. Sci. U.S.A. 98, 10757-10762
23. Rand, D. G., Ohtsuki, H., and Nowak, M. (2009) Direct reciprocity with costly punishment: Generous tit-for-tat prevails. J. Theor. Biol. 256, 45–57
24. Hilbe, C., and Sigmund, K. (2010) Incentives and opportunism: from the carrot to the stick. Proc. R. Soc. Lond. B. 277, 2427-2433
25. Chavez, A. K., and Bicchieri, C. (2013) Third-party sanctioning and compensation behavior: Findings from the ultimatum game. J. Econ. Psych. 39, 268–277
26. Schoenmakers, S et al. (2014) Sanctions as honest signals – The evolution of pool punishment by public sanctioning institutions. J. Theor. Biol. 356, 36-46
27. Nowak, M. A., and Sigmund, K. (1998) Evolution of indirect reciprocity by image scoring. Nature 393, 573–577
28. Milinski, M., Semmann, D., and Krambeck, H. (2002) Reputation helps solve the "tragedy of the commons." Nature 415, 424–426
29. Sylwester, K., and Roberts, G. (2013) Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. Evol. Hum. Behav. 34, 201–206
30. Santos, dos, M., Rankin, D. J., and Wedekind, C. (2011) The evolution of punishment through reputation. Proc. R. Soc. Lond. B. 278, 371–377
31. Roos, P. et al. (2014) High strength of ties and low mobility enable the evolution of third-party punishment. Proc. R. Soc. Lond. B. 281, 20132661

32. Santos, dos, M., Rankin, D. J., and Wedekind, C. (2013) Human cooperation based on punishment reputation. Evolution 67, 2446–2450
33. Fehr, E., and Fischbacher, U. (2003) The nature of human altruism. Nature, 425, 785–791
34. Kurzban, R., Descioli, P., and O'Brien, E. (2007) Audience effects on moralistic punishment. Evol. Hum. Behav. 28, 75–84
35. Piazza, J., and Bering, J. (2008) The effects of perceived anonymity on altruistic punishment. Evol. Psych. 6, 487-501
36. Rockenbach, B., and Milinski, M. (2011) To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. Proc. Natl. Acad. Sci. U.S.A. 108, 18307–18312
37. Barclay, P. (2006) Reputational benefits for altruistic punishment. Evol. Hum. Behav. 27, 325–344
38. Nelissen, R. (2008) The price you pay: cost-dependent reputation effects of altruistic punishment. Evol. Hum. Behav. 29, 242–248
39. Kiyonari, T., and Barclay, P. (2008) Cooperation in social dilemmas: free riding may be thwarted by second-order reward rather than by punishment. J. Pers. Soc. Psych. 95, 826–842
40. Horita, Y. (2010) Punishers May Be Chosen as Providers But Not as Recipients. Letters on Evolutionary Behav. Sci. 1, 6-9
41. Ozono, H., and Watabe, M. (2012) Reputational benefit of punishment: comparison among the punisher, rewarder, and non-sanctioner. Lett. Evol. Behav. Sci. 3, 21–24
42. Balafoutas, L., Nikiforakis, N. & Rockenbach, B. (2014) Direct and indirect punishment among strangers in the field. Proc. Natl. Acad. Sci. U.S.A. 111, 15924–15927
43. Gordon, D.S., Madden, J.R. & Lea, S.E.G. (2014) Both loved and feared: third-party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. Plos One 9, 11004540.
44. Brañas-Garza, P., Espìn, A.M., Exadaktylos, F. & Herrmann, B. (2014) Fair and unfair punishers co-exist in the Ultimatum Game. Sci. Rep. 4, 6025
45. Peysakhovic, A., Nowak, M.A. & Rand, D.G. (2014) Humans display a cooperative phenotype that is domain general and temporally stable. Nat. Comm. 5, ncomms5939
46. Espìn, A.M., Brañas-Garza, P., Herrmann, B. & Gamella, J.F. (2012) Patient and impatient punishers of free-riders. Proc. R. Soc. Lond. B. 279, 4923–4928
47. Carpenter, J. (2003) Is fairness used instrumentally? Evidence from sequential bargaining. J. Econ. Psychol. 24, 467-489

48. Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., Miura, A., Inukai, K., Takagishi, H. & Simunovic, D. (2012) Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. Proc. Natl. Acad. Sci. USA 109, 20364–20368
49. Fehr, E. & Rockenbach, B. (2003) Detrimental effects of sanctions on human altruism. Nature 422, 137-140
50. Xiao, E. (2013) Profit-seeking punishment corrupts norm obedience. Games Econ. Behav. 77, 321-344
51. Pedersen, E. J., Kurzban, R., and McCullough, M. E. (2013) Do humans really punish altruistically? A closer look. Proc. R. Soc. Lond. B. 280, 2012272342.
52. Lotz, S., Okimoto, T. G., Schlosser, T., and Fetchenhauer, D. (2011) Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. J. Exp. Soc. Psych. 47, 477–48043
53. FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J.J. & Phelps, E.A. (2014) Fairness violations elicit greater punishment on behalf of another than for oneself. Nat. Comm. ncomms6306
54. Selten, R. (1975) Reexamination of the perfectness concept for equilibrium points in extensive games. Int. J. Game Theor. 4, 25-55
55. Nowak, M. A., and Sigmund, K. (1992) Tit for tat in heterogeneous populations. Nature 355, 250–253

Box 1. The punisher's image score

Image scores associated with helping behaviours have typically been modelled as either positive or negative, where individuals with positive image scores are more likely to receive help from observers than individuals with negative image scores [27]. Individuals with an image score of zero are perceived neutrally by observers. We suggest that the image score associated with punishment depends on the emotions that observers experience towards the punisher and the context of the next interaction (Figure 1; Table 1). For ease, we have assumed that punishers can evoke the emotions of hatred, love or fear in observers, depending on the available information (though we note that this assumption requires empirical validation). Evoking the emotion of hatred is expected to be associated with negative image score; while evoking the emotion of love should have positive consequences for punishers. However, we suggest that evoking the emotion of fear in observers could be either beneficial or harmful to the punisher depending on whether observers are forced to interact with the punisher or can avoid them (Table 1).

Defecting punishers or individuals who punish cooperators might induce purely negative emotions like hate and disgust. Individuals who punish defectors could receive respect but also some fear as long as competitive motives underpinning punishment cannot be ruled out by observers (e.g. if the punisher was the victim of the cheat or will interact with the cheat in future). Although (justified) third-party punishment should usually be evaluated more positively than retaliatory punishment, even the moral legitimacy of third-party punishers will be called into question by observers if punishers stand to increase their payoff relative to that of the target (i.e. where punishment is efficient, [48, 50]). In contrast, 'inefficient' punishment (fee to fine ratio < 1, red arrows in Figure 1) should remove the doubt that punishment is motivated by a self-serving desire to improve relative payoffs. Under these circumstances, an individual could use justified punishment to signal reliably their willingness to uphold fairness norms and other-regarding preferences (e.g. [49]). Thus, inefficient punishers could achieve image scores as high as helpers, and the combination of inefficient punishment and helping might even yield the highest score, and hence evoke voluntary help in the absence of any punishment threats. This possibility needs both empirical and theoretical exploration, using experiments that dissociate reputation consequences of helping from reputation consequences of punishing. To date, most empirical studies on the effects of a punitive reputation have demonstrated that punishers induce fear in bystanders (e.g. [32, 33] but see [43]) rather than exploring conditions that maximise the chances that punishers induce love.

Glossary

Altruistic punishment: typically used to describe punishment that occurs in n-player games, such as the public goods game (see below). Punishment is described as altruistic because the punisher pays the cost of punishment while any benefits of increased within-group cooperation are shared among punishers and non-punishers. Note that punishment need not impose lifetime fitness costs on punishers and is therefore not necessarily altruistic in the true sense of the word.

Antisocial punishment: punishment that is aimed at individuals whose actions benefit, rather than harm, the group.

Cooperation: the outcome of a social interaction in which all players gain lifetime direct fitness benefits.

Public Goods Game: an n-player game where individuals make contributions to a communal venture. Collective benefits are greatest if everyone contributes to the resource but individuals do best to withhold investment and free-ride on the investments of others.

Punishment: the act of paying to reduce the payoff of another individual. There are many ways that a punisher might ultimately gain direct fitness benefits from this investment. Efficient punishment: fee to fine ratio > 1; inefficient punishment: fee to fine ratio ≤ 1.

Reputation: information about the previous behaviour of an individual that can be used to predict how they might behave in future.

Sanction: involves harming another individual but without incurring the cost involved in punishment.

Third-party punishment: typically refers to a scenario where a cheating individual is punished by an uninvolved bystander.

Trust Game: a two-player game where the Truster is endowed with a sum of money which they can entrust to the Trustee. Any money sent to the Trustee is multiplied by the experimenter and the Trustee can then choose how much of the endowment to send back to the Truster. Mutual benefits are highest if the Truster trusts the Trustee and the Trustee returns half the endowment to the Truster. However, Trustees gain higher payoffs by keeping any money endowed to them and Trusters are thus selected to not trust.

Volunteer's Dilemma: n-player Public Goods Game where the benefit of the public good is produced so long as one player chooses to cooperate. Benefits thus follow a non-linear increase with number of cooperators.

Outstanding questions

Are there conditions under which punishers might be rewarded for their actions by third-parties? If so, what are these conditions?

Under what circumstances, if any, are cooperative punishers preferred over cooperative non-punishers for interactions?

To what extent is justified punishment indicative of the punisher's future cooperative behaviour?

Figure 1: We propose that the reputation consequences of punishment depend on the punisher's role (active or passive) and behaviour in the game in which the reputation is built, on the type of game (2-player game, n-player game, 3rd party intervention) and on the fee to fine ratio (>1 or ≤1, indicated by black and red arrows, respectively). Depending on the combination of these factors a punisher might acquire a reputation for being antisocial, competitive or cooperative; and evoke hatred, fear or love in observers as a consequence.

Table 1: We predict how individuals will react to their co-players, depending on the co-player's punishment reputation, and on the strategic nature of the following interaction. To this end, we assume that individuals know their co-players' punishment reputation from the past. For the following interaction, we consider several scenarios, namely whether individuals can choose partners before the interaction takes place, the structure of the game itself (IR, IPD, IPGG), and whether individuals can punish each other after the interaction. If partners are chosen, we predict that only loved punishers are ever preferred as partners as much (or even more than) non punishing cooperators (NPC). The fear reputation of punishment is valid whenever players are forced to interact with one another and punishment remains an option in the game to be played. In the absence of punishment, we predict that only loved punishers receive voluntary help as much as NPC. Scenarios where acquiring a punitive reputation could could be under positive selection (assuming that the punisher's cooperativeness is directly inferred) are highlighted in red. Abbreviations: IR=Indirect Reciprocity Game, IPD=Iterated Prisoner's Dilemma, IPGG=Iterated Public Goods Game.

Game to be played	**Punisher's reputation**		
	Antisocial	**Competitive, potentially cooperative**	**Cooperative norm protector**
IR or IPD + partner choice + punishment	Avoided	Intermediate preference	Preferred over competitive partners; but ≤ NPC
IR or IPD + partner choice - punishment	Avoided	Intermediate preference	Preferred over competitive punishers; = NPC
IPGG + partner choice + punishment	Avoided	Intermediate preference or punishers might be preferred over NPC where defectors cannot be easily excluded from the game	
IPGG + partner choice - punishment	Avoided	Intermediate preference; but ≤ NPC	Preferred over competitive punishers; = NPC
IR - partner choice + punishment	No help	Punishers receive more help than NPC	
IR - partner choice - punishment	No help	No help	Help = NPC
IPD - partner choice + punishment	Defect	Cooperation received > NPC or provoke retaliation	Cooperation received > NPC
IPD - partner choice - punishment	Defect	Cooperation received ≤ NPC	Cooperation received = NPC
IPGG - partner choice + punishment	Depends what strategy other group members play	Cooperation received > NPC or provoke retaliation	Cooperation received > NPC
IPGG - partner choice - punishment	Levels of cooperation will be generally low		