CellPress

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

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. By contrast, if partner choice is possible and an individual's cooperative intent can be inferred directly, then individuals with a nonpunishing cooperative reputation should typically be preferred over punishing cooperators.

The puzzle of punishment

Punishment (see Glossary) 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, because 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 two-player and *n*-player games [2,6], although 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 coevolution between cooperating and punishing defectors [19–21], one has to ask under what conditions punishment could be favoured over alternative nondestructive or even value-building control mechanisms, such as reciprocal defection, rewarding cooperators, or compensating victims (e.g., [13,16,22–25]). Here, we are particularly concerned with peer punishment and ignore the evolution of centralised punishment institutions, although we note that similar arguments to those that we make about peer punishment might also apply to centralised punishment (e.g., [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. Given that 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, because punishers can then benefit from acquiring a punitive reputation. Reputation generally offers one major solution to the question of why

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 nonpunishers. Note that punishment need not impose lifetime fitness costs on punishers and, therefore, is 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 that 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, thus, trusters are 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. Thus, benefits follow a nonlinear increase with the number of cooperators.

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

0169-5347/
 http://dx.doi.org/10.1016/j.tree.2014.12.003

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 two-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 nonpunisher) 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 [21]. The latter would otherwise be expected to outperform responsible punishment strategies (e.g., [19,20]).

Empirical data on the reputation consequences of punishment are 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 nonpunishers (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]; 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 (N.J. Raihani and R. Bshary, unpublished, 2014). 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 in others. Here, we provide a conceptual framework (Figure 1; Box 1) that integrates existing theoretical and empirical knowledge when possible and 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 thereof, 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

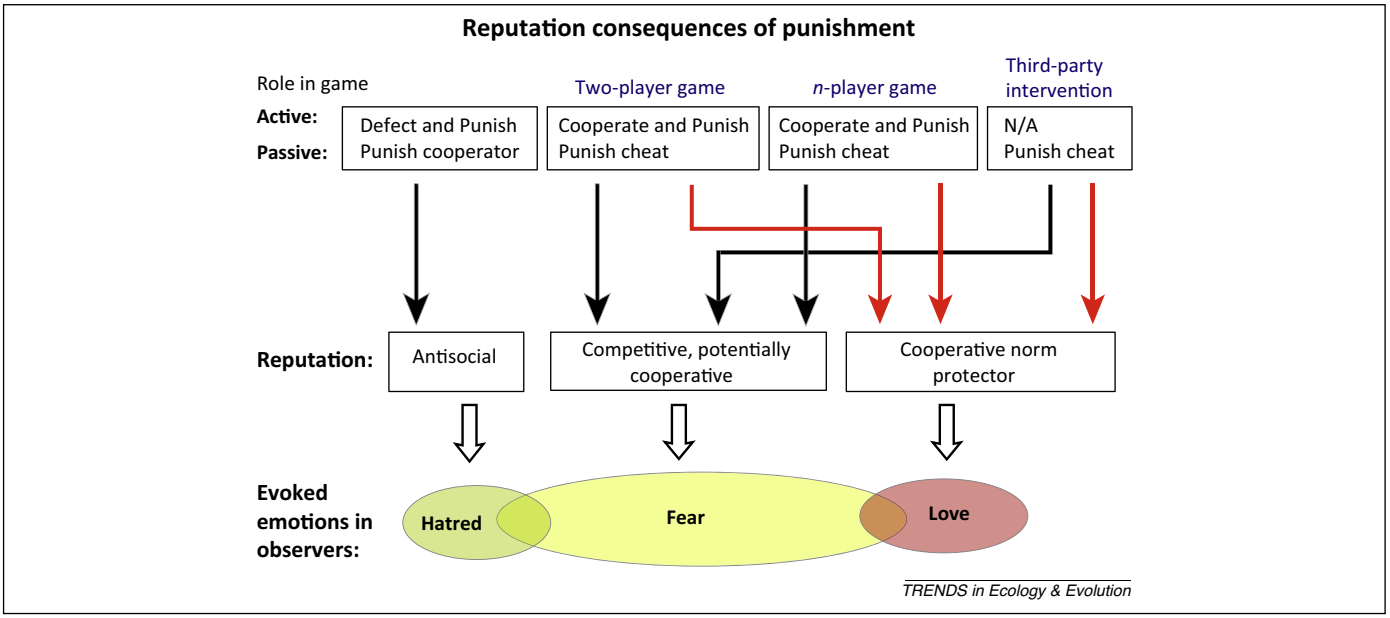

Figure 1. We propose that the reputation consequences of punishment depend on the punisher's role (active or passive) and behaviour in the game in which the reputation is built, on the type of game (two-player game, *n*-player game, or third-party intervention) and on the fee:fine (>1 or ≤1, indicated 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.