Review

Punishment and cooperation in nature

Nichola J. Raihani[1,2], Alex Thornton[3] and Redouan Bshary[4]

[1] Department of Genetics, Evolution and Environment, University College London, Gower St, London, WC1E 6BT, UK
[2] Institute of Zoology, Zoological Society London, Regent's Park, London, NW1 4RY, UK
[3] Department of Experimental Psychology, University of Cambridge, Downing St, Cambridge, CB2 EB, UK
[4] Université de Neuchâtel, UniMail, Institut de Biologie, Eco-Ethologie, Rue Emilie-Argand 11, CH-2000, Neuchâtel, Switzerland

Humans use punishment to promote cooperation in laboratory experiments but evidence that punishment plays a similar role in non-human animals is comparatively rare. In this article, we examine why this may be the case by reviewing evidence from both laboratory experiments on humans and ecologically relevant studies on non-human animals. Generally, punishment appears to be most probable if players differ in strength or strategic options. Although these conditions are common in nature, punishment (unlike other forms of aggression) involves immediate payoff reductions to both punisher and target, with net benefits to punishers contingent on cheats behaving more cooperatively in future interactions. In many cases, aggression yielding immediate benefits may suffice to deter cheats and might explain the relative scarcity of punishment in nature.

Punishment in nature: unresolved issues

Individuals are often tempted to cheat in social interactions, thereby gaining a benefit at the expense of cooperative partners. To encourage partners to behave cooperatively, individuals might therefore use control mechanisms that render cooperative behaviour a more profitable option than cheating for the partner. One such mechanism is punishment (see Glossary) [1]. Several laboratory studies have shown that punishment promotes cooperation among humans, typically using stylised laboratory games (e.g. [2–6]). By comparison, only a handful of studies have shown that punishment promotes cooperation among non-human animals [7–11]. This relative paucity of evidence prompted arguments about why initial predictions that punishment should be common [1] do not fit current data [12,13].

Here, we critically assess empirical evidence for punishment in non-human species. We first outline how punishment can be distinguished from other forms of aggression that promote cooperative behaviour, such as coercion and sanctions, and then go on to discuss specific empirical examples of punishment. We end by discussing the conditions that are likely to favour punishment over alternative control mechanisms and whether these conditions are likely to be met in non-human species.

What is (and what is not) punishment?

Following the seminal paper by Clutton-Brock and Parker [1], we assert that punishment occurs when an individual reduces its own current payoffs to harm a cheating partner. In doing so, the punisher reduces the payoffs of the cheat and thereby promotes cooperative behaviour from the cheat in subsequent interactions (Box 1). Thus, punishment is equivalent to 'negative reciprocity' [1,14]. This functional definition is useful for studying punishment among non-human animals because punishment is not contingent on a capacity for mental state attribution and does not require the punisher to be aware of how its behaviour might influence that of the target [1]. Punishers need not always be involved in the initial interaction with the cheat. For example, in 'policing' or 'third party punishment', a bystander observes a cheat and is willing to reduce its own current payoffs to reduce the payoff to the cheat. It is still largely unclear how punishers benefit from third-party punishment, however [14–18].

As with punishment, some other control mechanisms also rely on responses to cheating that reduce the payoffs to cheats. However, unlike punishment, such responses do not necessarily reduce the current payoffs of the actor. Instead, several responses to cheating described in the literature are immediately self-serving and, hence, do not rely on future benefits arising from the increased cooperative behaviour of the target to be under positive selection. These examples do not fit the negative reciprocity concept but are

Glossary

Centralised punishment: punishment devolved to a legitimate authority.
Coercion: occurs when one player is forced into interacting with and cooperating with an aggressive partner. Under coercion, the coerced individual would do better to terminate the interaction but is somehow prevented from doing so.
Negative pseudo-reciprocity: occurs when a cheating behaviour by one individual allows the partner to perform a self-serving response, which harms the cheating individual as a byproduct.
Negative reciprocity: see punishment.
Peer punishment: punishment carried out by other members of the social group of a cheat.
Prisoner's dilemma games: two-player games in which players have the option to either cooperate or defect. In a one-shot game, players receive the highest payoff from defecting, regardless of the behaviour of the partner. However, mutual payoffs are highest when both players cooperate.
Public goods games (PGGs): players are endowed with an initial sum of money, some or all of which they may contribute to the communal pot. Contributions to the communal pot are increased by the experimenter and then divided among all players in the game, regardless of who contributed. In these games, the most profitable strategy is to withhold contributions and 'free-ride' on the investments of others.
Punishment: occurs when an individual reduces its own current payoffs to harm a cheating partner. In doing so, the punisher reduces the payoffs to the cheat and thereby promotes cooperative behaviour from the cheat in subsequent interactions.
Sanctions: one form of negative pseudoreciprocity. Sanctions occur when two or more players interact and one player cheats by withholding investment. The cheated partner then performs a self-serving behaviour to terminate the interaction, which harms the cheating partner as a byproduct.

Corresponding author: Raihani, N.J. (nicholaraihani@gmail.com).

 0169-5347/$ – see front matter doi:10.1016/j.tree.2011.12.004 Trends in Ecology and Evolution, May 2012, Vol. 27, No. 5

Box 1. Payoffs associated with punishment and sanctions

Figure I shows categories of interactions among non-kin. The arrows indicate the initial instigator and recipient and the + or − signs represent fitness changes (as in [1]).

Punishment

In punishment (Figure Ia), one individual cheats a partner, thereby increasing its immediate payoffs (++ relative to the payoff increase associated with cooperating, +) and imposing a fitness cost, −, on the partner. The partner then retaliates with a behaviour that reduces its immediate payoffs further, −, which imposes a fitness cost on the cheat. The fitness costs experienced by the cheat can be equal to (−) or greater than (−−) the payoff losses to the punisher of executing the punishment. In response to punishment, the cheat behaves more cooperatively in subsequent interactions with the punisher. Note that the cooperative interaction is mutually beneficial (in fitness terms) to both players compared to outside options (not interacting) but that, in the absence of punishment, the cheat could gain higher payoffs from exploiting a cooperative partner. Also note that some effects are on lifetime fitness (the effects of being cheated, of being punished and of mutual cooperation), whereas others are on immediate payoffs (the effects of cheating and of punishing). In interactions where cooperative behaviour is binary, the cheating is simply the opposite of cooperative behaviour. In interactions where cooperative behaviour can be a continuous investment, then cheating can be defined as any investment that is less than the population mean [55].

Sanctions

Sanctions (Figure Ib) occur when one individual cheats a partner, thereby gaining a payoff increase relative to cooperating [56]. In response, the partner performs a self-serving behaviour that imposes costs on the cheat as a byproduct (2). This self-serving behaviour also serves to end the interaction. Thus, sanctions fit the concept of negative pseudo-reciprocity. We note that the term 'sanction' has also been used differently in the literature (e.g. [57–59]). Under the concept of sanctions, there is no future to the interaction. Classic examples include the selective abortion by yucca trees of fruits that harbour too many seed-eating larvae of its pollinator, the yucca moth [60]; and the selective inhibition of nodule growth by leguminous plants in root parts in which rhizobia partner bacteria fail to fix significant amounts of nitrogen [61]. The permanent eviction of uncooperative individuals from a territory or group also fits the sanctions concept. For example, in coral-dwelling gobies (*Paragobiodon xanthosomus*; Figure II) dominant individuals sometimes evict similar-sized subordinates, because subordinates that grow too large can threaten the superior status of their dominant neighbour [62]. Dominant gobies benefit from evicting overgrown competitors without the need for subordinates to behave more cooperatively (by reducing growth rate) in future. Indeed, evictees rarely return to their group and there is therefore little or no potential for them to cooperate more in response to eviction from dominants. Noë [63] pointed out that sanctions often occur within a biological market in which individuals choose the best partner out of a possible range. Although the simple threat of terminating an interaction can be enough to promote cooperative behaviour [64], the additional opportunity of partner switching could enhance the effect [65–67]. Indeed, switching to a different partner is an efficient way to select against cheating in marine cleaning mutualisms [68].

Figure II. A coral-dwelling goby. Dominant gobies sanction subordinates that breach a defined size threshold by evicting them from the group. Reproduced, with permission, from Joao Paulo Krajewski.

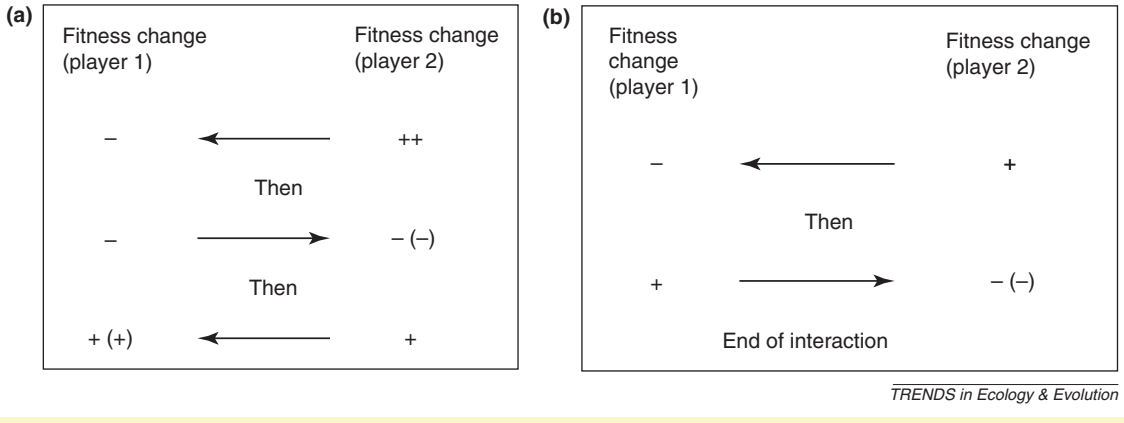

Figure I. Categories of interactions among non-kin. **(a)** punishment and **(b)** sanctions.

instead cases of sanctions or 'negative pseudo-reciprocity' [19,20] (Box 1).

Coercion is another form of aggressive behaviour that can induce cooperative behaviour in the target. However, coercion differs from punishment because, from the point of view of the target, no interaction yields a higher payoff than interacting and cooperating with the aggressor. Thus, coercion occurs when targets of aggressive behaviour would do best to avoid interactions with the aggressor but are somehow prevented from exercising this higher paying outside option. Experiments on a coordination task in keas (*Nestor notabilis*) provide a good example [21]. In the experiment, one individual had to sit on a lever to lift a lid covering a food tray, thereby allowing another individual to feed. Under these conditions, dominant birds aggressively forced subordinate partners to sit on the lever

Box 2. The evolution of punishment in *n*-player games

Experimental studies investigating the evolution of cooperation in *n*-player games have typically used *n*-player prisoner's dilemma (NPD) payoffs, rendering contributions altruistic [69,70] and resulting in the tragedy of the commons [71]. Evidence indicates that humans willingly punish free-riders and that targets subsequently behave more cooperatively [36–40]. However, in one-shot games, this raises a second-order social dilemma as punishers invest in harming free-riders although the resulting benefit (of increased cooperation) is shared among punishers and non-punishers alike. Nevertheless, because punishment often promotes cooperative behaviour in one-shot games, evolutionary explanations for its emergence and stability have been proposed. For example, several authors [14,72,73] have suggested that punishment could spread through cultural group selection. Here, social learning facilitates the local spread of punitive behaviour, and demes with a high number of punishers outperform demes without punishment. Arguments over the importance of cultural group selection for the evolution of punishment have centred on two issues. First, Gardner and West [15] and Lehmann *et al.* [16] pointed out that, because punishment is altruistic in these models, it relies on kin selection to spread in a population. Therefore, the logic of inclusive fitness theory still applies. Second, cultural group selection models struggle to explain how punishment becomes established when it is initially rare [15–17]. Others have argued that punishment in one-shot games occurs because humans evolved in a social system in which repeated interactions are the norm and often take place in a communication network [74]. Thus, humans are error-prone when confronted with anonymous one-shot interactions [18,75–78].

More generally, explanations for the evolution of punishment might have been hindered by the use of inappropriate payoff matrices. Specifically, punishment is efficient when it reduces a free-rider's payoff below the population average; at this point, the target does best to contribute rather than free-ride. Further punishment is wasteful because it reduces group gains without increasing contributions from targets. Therefore, the net benefits of punishment are not a linear function of contributions as assumed in the NPD framework but are better described with a step function as assumed in the volunteer's dilemma game [79,80]; Figure I). In non-linear public goods games, cooperators and punishers are expected to coexist in a stable mixed equilibrium [81,82] and punishers can also invade when they are initially rare.

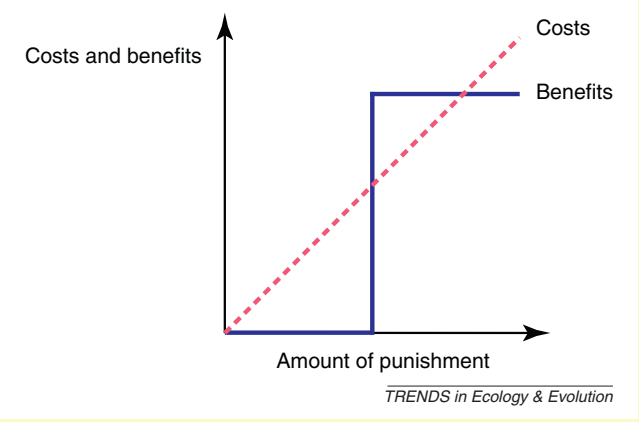

Figure I. Simple schematic of the costs and benefits of punishment according to the investment in punishment under the volunteer's dilemma framework. Although the costs increase linearly with increasing investment in punishment, the benefits (in terms of increased future cooperative behaviour from the punished individual) follow a step function. Any investment below the threshold yields no benefits, whereas further investment above the threshold yields no additional benefits. Reproduced, with permission, from [80].

without ever reciprocating. Forced copulations in animals also fit the concept of coercion [22].

Mixed evidence that punishment promotes cooperation

The effect of punishment on cooperation has been best studied in humans, typically using *n*-player public goods games (PGGs) under controlled laboratory settings [2–5,23]. Players can punish free-riders by paying a small fee to impose a larger fine on the cheat. Although there is still some debate surrounding the evolutionary scenarios (Box 2), the majority of studies have shown that punishment promotes cooperation in *n*-player games (reviewed in [23–26]), although considerable cross-cultural differences in the administration and effects of punishment exist [27,28]. Punishment might be less successful in two-player interactions. Dreber *et al.* [6] used iterated two-player prisoner's dilemma games with and without a punishment option and found that, although punishment promoted cooperative behaviour, punishers achieved lower payoffs than did non-punishers. Instead, players that responded to cheats with reciprocal defection achieved the highest payoffs. In this experiment, punishers were disadvantaged by the relatively short time horizon of expected interactions with the current partner. Interactions lasted between one and nine rounds, which according to another recent experimental study [5], is insufficient for punishers to recoup their initial investment in harming a cheating partner.

In contrast to the large number of laboratory studies, there have been relatively few real-world studies of punishment and cooperation in humans. Notable exceptions have focused on hunter-gatherer societies and typically describe centralised punishment rather than peer punishment (see Glossary) [29–32]. None of these studies have explicitly examined whether punishment causes an increase in the future cooperative behaviour of the target. Thus, these field studies do not help to elucidate the precise conditions that would favour punishment over alternative control mechanisms, such as terminating the interaction with a cheating partner, partner switching, or responding with reciprocal cheating (but see [33] for a theoretical approach). Similarly, very little work has addressed questions about the form that punishment is likely to take in reality and about the relative efficacy of different types of punishment. For example, rather than monetary fines, punishment can also take the form of physical aggression, verbal reprimands, negative gossip statements or ostracism [29,34–36]. These different types of punishment might impose variable costs on cheats and differentially affect their propensity to cooperate. More data on punishment in humans under real-world settings are clearly a research priority.

In contrast to human laboratory studies, relatively few studies have demonstrated experimentally that non-human animals use punishment to promote cooperation. Perhaps surprisingly, evidence for punishment in closely related non-human primate species is scarce. For example, although a capacity for vengeful behaviour has been demonstrated under laboratory conditions in chimpanzees *(Pan troglodytes)* [37], there is very little evidence that individuals punish cheats under real-world settings (e.g. for failure to reciprocate grooming or provide support,

[38]). Moreover, in the laboratory study, vengeful behaviour in response to food theft by conspecifics decreased over time at the same time that thefts increased [37]. This further argues against the idea that vengeful behaviour in chimpanzees functions as a form of punishment, at least in this experimental context. Aggression from dominant rhesus macaques *(Macaca mulatta)* towards subordinates that fail to advertise food patches vocally [39] also superficially resembles punishment. However, a plausible alternative explanation is that failure to claim possession of food patches vocally results in resource-based conflict among the monkeys [40]. In addition, there is no evidence to suggest that dominant aggression increases the chances that subordinates will advertise food patches in future; indeed, aggression in this context might be fundamentally unlikely to promote cooperative food-calling behaviour (Box 3).

Solid evidence of punishment among non-human species has come from work on the mutualism between bluestreak cleaner wrasse (*Labroides dimidiatus*) and their reef-fish clients. Observations conducted under natural conditions have shown that clients often aggressively chase cleaners after jolting. Jolts are a correlate of mucus feeding by cleaners, which constitutes cheating. Following punishment, jolt rate subsequently declines [7]. Experimentally preventing clients from punishing cleaners (by anaesthetising them), showed that punishment causally promotes cooperative behaviour from cleaner fish [7]. Client aggression in response to cleaner cheating fits the concept of punishment because the clients experience an initial reduction in payoffs when they chase cheating cleaners. Chasing is not necessary to terminate an interaction; some client species simply swim off instead [7]. The investment in punishment is repaid if this cleaner subsequently provides a better cleaning service (i.e. more ectoparasite removal with less biting). Punishment also promotes cooperation within mixed-sex pairs of cleaner fish, with male–female pairs occasionally working together to clean a joint client fish (Figure 1). During pairwise inspections, the male and female cleaner fish face a problem akin to a prisoner's dilemma because only one of the pair can obtain the benefit of biting the client whereas the cost (of client departure) will be experienced by both cleaners, regardless of who cheated [41]. Despite the apparent temptation to cheat before the partner does, pairs of cleaner fish provide a better cleaning service than do singletons [41]. The improved service quality is almost entirely the result of increased cooperative behaviour by females. Males aggressively punish females that cheat during joint inspections of model clients and this incentivises females to feed more against their preference in subsequent inspections with that male [10]. Client aggression towards cheating cleaners and male aggression towards cheating females fit the concept of punishment rather than coercion. This is because, under natural circumstances, cleaners choose to interact with clients and

Box 3. Learning processes and endocrine mechanisms of punishment

Effective punishment requires some means by which aggressive behaviour from punishers causes victims to behave more cooperatively in future. A possible mechanism is operant conditioning, whereby individuals learn to associate their behaviour with a particular outcome [83]. If a particular behaviour is reliably followed by an aversive consequence, the animal should learn not to perform the behaviour in future. By contrast, aggression is unlikely to induce learning if it is aimed at targets that failed to do something unless the range of possible responses to the punishment is tightly constrained. For example, consider the pay-to-stay hypothesis, according to which helpers in cooperatively breeding societies contribute to cooperative activities in return for being allowed to reside on the territory, and breeders attack 'lazy' helpers (Figure I) [84]. Despite suggestive evidence of punishment for failure to help (e.g. [85–87]), no study has shown conclusively that aggression causes lazy helpers to increase their contributions to cooperative activities. Under natural conditions, the range of possible behaviours a lazy helper could be performing at any given time is vast. A lazy helper is therefore highly unlikely to learn to respond to breeder aggression by increasing its investment in cooperation, rather than simply learning to avoid the attacker. Indeed, in the absence of language to explain the rationale for punishment, it is difficult to envisage how an individual that is punished for omissions could learn to behave appropriately in future. This learning criterion might also mean that punishment is unlikely to enforce cooperative food-calling behaviour in rhesus macaques, where subordinates are often attacked for failure to advertise food patches [52].

Rather than relying purely on learning processes, punishment might also operate through endocrine mechanisms. Aggression might often cause a stress response in the target [88]. If stress causes more cooperative behaviour then aggression functions as punishment. Few studies have addressed the link between stress and levels of cooperation. In meerkats (*Suricata suricatta*), high levels of cortisol in helpers are associated with elevated contributions to feeding the pups of the dominant pair [89]. However, escalating aggression typically leads to the temporary eviction of subordinate females from group, resulting in stress-induced abortion [90]. Dominant aggression therefore seems to be used primarily to suppress subordinate reproduction, rather than to promote contributions to cooperative activities. Cleaner wrasses behave more cooperatively after being exposed to a stressor [91]. However, punishment by a client appears to induce increased cooperation during the next interaction between cleaner and punisher rather than a general increase in cooperative behaviour towards all clients. It therefore remains to be seen whether hormonal mechanisms could have sufficiently specific effects to serve as the basis through which punishment induces cooperation.

TRENDS in Ecology & Evolution

Figure I. Naked mole-rat queen and workers. Evidence that dominant breeders use punishment to activate lazy workers in cooperatively breeding species is currently scarce, and punishment may be fundamentally unlikely to evolve in this context.

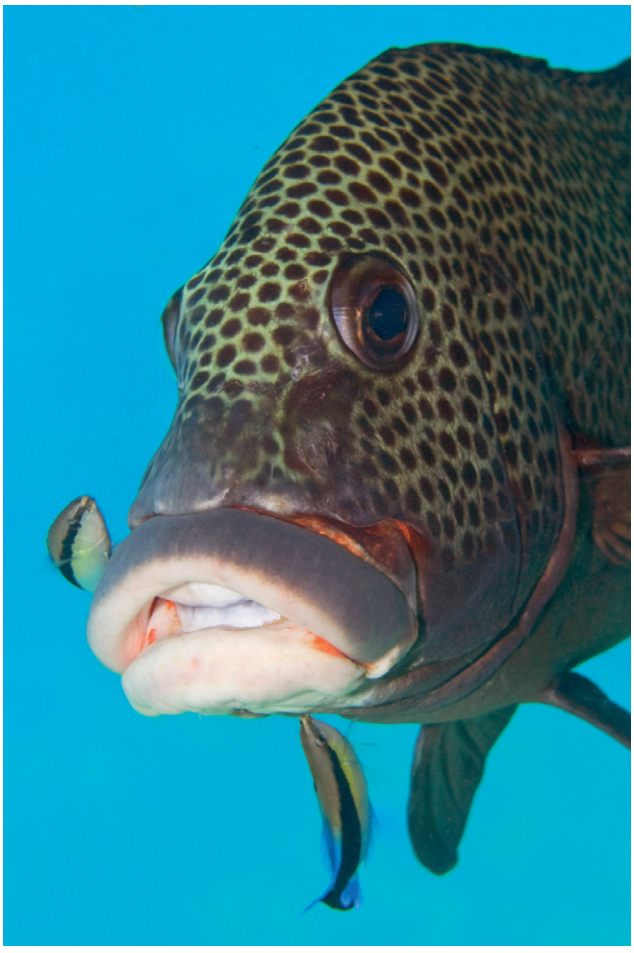

Figure 1. A male–female cleaner fish pair working together to clean a joint client. Male punishment in response to cheating females causes females to feed more against their preference in subsequent interactions. Reproduced, with permission, from Joao Paulo Krajewski.

females choose to perform joint inspections of clients with males, rather than inspecting clients alone. Thus, it can be assumed that, in both cases, the punished individual experiences higher payoffs from continuing the interaction than from pursuing an outside option of not interacting with the punisher, even if punishment imposes strategic constraints on the behaviour of the target.

Punishment also occurs in interactions between sabre-tooth blennies *(Plagiotremus* sp.) and their victims [9]. Blennies are parasitic fish that feed by opportunistically sneaking up to victims and removing scales or mucus [42]. Recent work using a common victim species of the blennies, the scalefin anthia (*Pseudanthias squamipinnis*), showed that aggressive chasing of blennies by anthias reduces the probability that the blenny will target anthias for its next attack. Thus, in this case, punishment by individual anthias creates a public good for the shoal because the blenny is then less likely to attack the punisher and the punisher's shoal members. Nevertheless, punishment in this case may still benefit the individual punisher. Experimental work with model targets in the lab demonstrated that blennies can distinguish between punishers and look-alike non-punishers, and avoid biting targets that punish.

To sum up, good evidence from controlled laboratory studies exists to show that humans use punishment to promote cooperation. Similar real-world studies in humans are relatively scarce, however. In non-human animals, studies on just a handful of species have been able to demonstrate that individuals invest to harm cheating partners. The studies in non-human animals that do demonstrate punishment have all done so under ecologically valid conditions.

The evolution of punishment in nature

By comparing empirical results from studies on humans and other species, it is possible to generate predictions about the game theoretic conditions that are likely to favour the evolution of punishment in nature. In turn, we argue that it is crucial to consider the following attributes of an interaction to predict whether punishment is likely to evolve as a cooperation-enforcing mechanism: (i) player number and asymmetries; (ii) time horizons for interactions; (iii) whether punished individuals are likely to learn the association between cheating and receiving aggression; and (iv) whether cheating is a continuous or a discrete event.

Player number and asymmetries

Punishment might generally be more likely to occur in symmetric *n*-player games than in symmetric two-player games. In two-player games where players have equal strength and strategic options, reciprocal defection might be a cheaper way to control cheating partners than punishment because the latter reduces current payoffs and may precipitate counter-punishment (e.g. [6]). The threat of retaliation increases the costs associated with punishment beyond the initial investment in harming a cheating partner. Reciprocal defection is less effective in *n*-player games, however, because defection harms cooperative partners as well as cheats. Thus, punishment directed at cheats might be more effective at promoting cooperation and, therefore, more likely to evolve, in *n*-player interactions. Nevertheless, punishment might be relatively common in two-player games where there are substantial asymmetries between players, such as differences in strength or strategic options. Differences in strength reduce the chance that punished individuals will retaliate rather than cooperate and, therefore, reduce the costs associated with punishment (e.g. [6,10]). As initially pointed out [1], punishment might be most likely to evolve under this 'common-sense' scenario: dominants are most often expected to punish subordinates that, in turn, are unlikely to retaliate. Asymmetries in player strength are expected to be a common feature of real-world interactions. Real-world studies of humans have shown that punishment is often administered by a central authority [30,43,44]. The inherent power asymmetry in such interactions might decrease the chances that punished individuals will retaliate against punishers. Where punishment is not institutionalised and is instead effected by the peers of the cheating individual (e.g. [29,31]), punishment might only occur if sufficient peers agree to participate in the punishing behaviour. For example, warriors of the pastoralist Turkana society from East Africa, often mete out corporal

punishment to free-riders, that is, individuals that defect during raids on other societies. Punishment is only administered once a critical number of warriors are assembled to admonish the cheat [31]. This strength in numbers can create an asymmetry between punishing and punished individuals, and reduce the possibility that punished individuals will retaliate against their aggressors. As well as asymmetries in strength, it might often be the case that players have asymmetric strategy sets, for example where only one class of player can cheat whereas the partner cannot. For example, cleaner fish can cheat non-predatory clients by biting them but these clients cannot cheat in return. Where one class of player cannot use reciprocal defection to control cheating partners, they might instead use punishment to induce cooperative behaviour (e.g. [8,10]).

Time horizons for interactions

The number of rounds in which the punisher and punished individuals expect to interact will influence the net benefit associated with investing in punishment. Short time horizons mean that there are fewer opportunities for the punisher to benefit from changes in the behaviour of the punished individual; whereas the net benefit of investing in punishment might be higher in more stable relationships (e.g. [5]). Indeed, peer-punishment in humans is thought to have evolved in small, stable groups, where punishers might have expected to interact with punished individuals again in the future. In larger groups with infrequent interactions, decentralised peer-punishment is expected to be less common and punishment responsibilities are instead devolved to centralised authorities [43].

The outside options available to interaction partners can affect the time horizons of interactions. Individuals pursue outside options when they leave, evict or eliminate a current partner rather than continuing with the current interaction [45]. Exercising outside options therefore necessarily shortens the time horizon of interactions. If individuals derive greater payoffs from exercising outside options rather than administering punishment or cooperating in response to punishment, respectively, then punishment is unlikely to evolve as a cooperation-enforcing mechanism. In many cases, individuals exercise outside options by terminating a current interaction with a cheating partner and instead seeking interactions with alternative partners. The payoffs of exercising partner choice, rather than punishment, to control interaction partners is likely to depend on the costs associated with finding a new partner and on population-level variance in cooperative tendency. Specifically, partner choice might be relatively common where the opportunity or energetic costs associated with finding a new partner are sufficiently low and also if there is sufficient population-level variation in partner cooperativeness such that new partners are likely to be more cooperative than the current (cheating) partner [46]. Field data on *L. dimidiatus* support the idea that outside options influence the time horizon of interactions and thereby affect the efficacy of punishment as a cooperation-enforcing mechanism. Only individuals of species that are forced to interact repeatedly with a specific cleaner fish (owing to small home range sizes) punish cheating cleaners, whereas clients with access to several cleaners (outside options) simply choose another cleaner fish if they receive a poor service. Conversely, the closely related cleaner wrasse *Labroides bicolor*, which roves over large areas [47], behaves more cooperatively in areas it visits frequently than in areas it rarely visits, as predicted by the 'shadow of the future' concept [48].

Punishment might often require associative learning

The relative paucity of data fitting the concept of punishment in non-humans might partly stem from cognitive constraints. Punishment requires that the punished individual associates its cheating behaviour with a negative response from the partner, and so learns to avoid repeating that behaviour again in subsequent interactions. It is currently unclear how such learning mechanisms might operate and whether different contexts might favour or preclude learning to cooperate in response to being punished (Box 3). This learning requirement can help researchers to predict more accurately the circumstances that are likely to favour punishment in nature. Animals are most likely to form associations between their own behaviour and the punitive responses of a partner when the punishment occurs very shortly after their own cheating behaviour. Long time delays between cheating behaviour and punishment are not conducive to associative learning [49,50]. Furthermore, individuals might be more likely to learn to stop performing a cheating behaviour than to learn to start performing a beneficial behaviour in response to being punished (Box 3).

It is also worth considering the cognitive mechanisms that might favour investment in punishment. Punishers might be relatively unlikely to learn a positive association between punishment and increased payoffs because punishment involves an immediate payoff reduction to punishers, compared with not punishing a cheat [12]. Instead, punishment might rely on fixed responses to cheats, on the capacity to predict that aggression will cause targets to behave more cooperatively, or on evolved subjective reward mechanisms (as in humans, where punishment activates areas associated with processing rewards in the brain [51]). In situations where individuals interact with several partners in different contexts, it could also be cognitively demanding for would-be punishers to keep track of who does what and when. This in turn might limit the capacity for punishment, in much the same way that it has been argued that similar cognitive burdens might inhibit the evolution of cooperation by positive reciprocity [52].

Cheating: discrete or continuous?

Punishment will be most likely to occur where cheating is a discrete event, rather than a continuous behaviour where the magnitude of inflicted costs is a function of interaction duration. For example, Turkana warriors sometimes desert their fellow fighters during a raid. This is a discrete cheating behaviour that often results in punishment [31]. Where cheating is a continuous behaviour, individuals can control cheating partners using aggression resulting in immediate benefits. This violates a key feature of the definition of punishment that posits an immediate payoff reduction and delayed benefits that are contingent on the

future behaviour of the target. For example, elephant seal *(Mirounga leonina)* pups that are caught suckling from a female other than their mother risk being attacked by this female, sometimes fatally [53]. Although such attacks might serve to deter pups from suckling from that female in the future, the female in question immediately benefits when she prevents a foreign pup from draining her milk resources. Thus, this does not fit the concept of punishment. By contrast, a cleaner fish that bites a client performs a discrete event that causes harm to the client. Clients are aggressive after the bite has occurred, meaning that aggression does not produce immediate benefits by stopping the cleaner fish from continuing to bite. Rather, the benefits of client aggression are conditional and depend on the future cooperative behaviour of the cleaner fish in question. Considering the difference between aggression that provides immediate benefits and aggression where the benefits are delayed (punishment) is not trivial. In the former situation, the benefits to the aggressor are assured as it performs the aggressive behaviour, whereas in the latter situation, the benefits are contingent on future interactions with the cheating partner (see [54] for a discussion of assured versus conditional benefits of investment).

Concluding remarks

Punishment is most likely to evolve in response to free-riding in symmetric *n*-player public goods games or in asymmetric two-player interactions. Although asymmetric two-player games and *n*-player public goods games are relatively common in nature, punishment has only rarely been documented in non-human species. This is in stark contrast to human behaviour in laboratory experiments. The scarcity of punishment among non-human animals might stem in part from cognitive constraints on what can be learned in response to aggression in the absence of language. Furthermore, it might often be the case that aggression against cheating partners yields immediate benefits, whereas immediate payoff reduction (and hence punishment) is only expected if the aggression follows discrete cheating events. In humans, laboratory games have all assumed that cheating is a discrete event, which could explain why games using human subjects report relatively high levels of punishment. Further 'field' data on human punishment would allow the assessment of whether cheating is often discrete or continuous and how commonly punishment is used to promote cooperation, compared to other mechanisms of interest. Finally, we note that there might be several examples where an aggressive act in response to a cheat appears to yield both immediate benefits to the aggressor (by stopping the cheat from performing a harmful behaviour) and future benefits (by making it less probable that the target will cheat in future interactions with the aggressor). Such interactions do not fit either our definition of punishment or of sanctions. We also note that, in some cases, an aggressive act could not only immediately benefit the actor, but also result in future, population-level benefits by causing the target to behave more cooperatively with future interaction partners. Thus, under our current definition, sanctioning behaviour could theoretically provide population-level public goods in the same way that punishment can [9].

Acknowledgements

We thank Marco Archetti, Mike Cant, Joah Madden, Manfred Milinski and Michael Taborsky for useful discussions about punishment. Thanks also to Chris Faulkes and Marion Wong for help finding images. The manuscript was greatly improved by the comments of three anonymous referees and Paul Craze. NR is funded by a Royal Society University Research Fellowship; AT is funded by a BBSRC David Phillips Research Fellowship; and RB is funded by the Swiss National Science Foundation.

References

1 Clutton-Brock, T.H. and Parker, G.A. (1995) Punishment in animal societies. *Nature* 373, 209–216
2 Fehr, E. and Gächter, S. (2000) Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* 90, 980–994
3 Fehr, E. and Gächter, S. (2002) Altruistic punishment in humans. *Nature* 415, 137–140
4 Rockenbach, B. and Milinski, M. (2006) The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444, 718–723
5 Gächter, S. *et al.* (2008) The long-run benefits of punishment. *Science* 322, 1510
6 Dreber, A. *et al.* (2008) Winners don't punish. *Nature* 452, 348–351
7 Bshary, R. and Grutter, A.S. (2002) Asymmetric cheating opportunities and partner control in a cleaner fish mutualism. *Anim. Behav.* 63, 547–555
8 Bshary, R. and Grutter, A.S. (2005) Punishment and partner switching cause cooperative behaviour in a cleaning mutualism. *Biol. Lett.* 1, 396–399
9 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
10 Raihani, N.J. *et al.* (2010) Punishers benefit from third-party punishment in fish. *Science* 327, 171
11 Raihani, N.J. *et al.* (2012) Male cleaner wrasses adjust punishment of female partners according to the stakes. *Proc. R. Soc. Lond. B* 279, 365–370
12 Brosnan, S.F. *et al.* (2010) The interplay of cognition and cooperation. *Philos. Trans. R. Soc. Lond. B* 365, 2699–2710
13 Bshary, R. and Bronstein, J.L. (2011) A general scheme to predict partner control mechanisms in pairwise cooperative interactions between unrelated individuals. *Ethology* 117, 1–13
14 Boyd, R. *et al.* (2003) The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. U.S.A.* 100, 3531–3535
15 Gardner, A. and West, S.A. (2004) Cooperation and punishment, especially in humans. *Am. Nat.* 164, 753–764
16 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
17 Fowler, J.H. (2005) Altruistic punishment and the origin of cooperation. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7047–7049
18 dos Santos, M. *et al.* (2011) The evolution of punishment through reputation. *Proc. R. Soc. Lond. B* 278, 371–377
19 Bergmüller, R. *et al.* (2007) Integrating cooperative breeding into theoretical concepts of cooperation. *Behav. Process.* 76, 61–72
20 Bshary, R. and Bergmüller, R. (2008) Distinguishing four fundamental approaches to the evolution of helping. *J. Evol. Biol.* 21, 405–420
21 Tebbich, S. *et al.* (1996) Social manipulation causes cooperation in keas. *Anim. Behav.* 52, 1–10
22 Clutton-Brock, T.H. (1995) Sexual coercion in animal societies. *Anim. Behav.* 49, 1345–1365
23 Gächter, S. and Herrmann, B. (2009) Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philos. Trans. R. Soc. Lond. B* 365, 2619–2626
24 Chaudhuri, A. (2010) Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Exp. Econ.* 14, 47–83
25 Balliet, D. *et al.* (2011) Reward, punishment, and cooperation: a meta-analysis. *Psychol. Bull.* 137, 594–615
26 Milinski, M. and Rockenbach, B. (2011) On the interaction of the stick and the carrot in social dilemmas. *J. Theor. Biol.* DOI: 10.1016/J.JTBI.2011.03.014
27 Rand, D.G. *et al.* (2009) Positive interactions promote public cooperation. *Science* 325, 1272–1275
28 Henrich, J. *et al.* (2006) Costly punishment across human societies. *Science* 312, 1767–1770

29 Wiessner, P. (2005) Norm enforcement among the Ju/'hoansi bushmen: a case of strong reciprocity? *Hum. Nat.* 16, 115–145
30 Rustagi, D. *et al.* (2010) Conditional cooperation and costly monitoring explain success in forest commons management. *Science* 330, 961–965
31 Mathew, S. and Boyd, R. (2011) Punishment sustains large-scale cooperation in prestate warfare. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11375–11380
32 Baumard, N. and Liénard, P. (2011) Second- or third-party punishment? When self-interest hides behind apparent functional interventions. *Proc. Natl. Acad. Sci. U.S.A.* 108, E753
33 Hilbe, C. and Sigmund, K. (2010) Incentives and opportunism: from the carrot to the stick. *Proc. R. Soc. Lond. B* 277, 2427–2433
34 Masclet, D. *et al.* (2003) Monetary and nonmonetary punishment in the voluntary contributions mechanism. *Am. Econ. Rev.* 93, 366–380
35 Carpenter, J.P. *et al.* (2004) Cooperation, trust, and social capital in Southeast Asian urban slums. *J. Econ. Behav. Organ.* 55, 533–551
36 Cinyabugama, M. *et al.* (2005) Cooperation under the threat of expulsion in a public goods experiment. *J. Pub. Econ.* 89, 1421–1435
37 Jensen, K. *et al.* (2007) Chimpanzees are vengeful but not spiteful. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13046–13050
38 Koyama, N.F. *et al.* (2006) Interchange of grooming and agonistic support in chimpanzees. *Int. J. Primat.* 27, 1293–1309
39 Hauser, M.D. (1992) Costs of deception: cheaters are punished in rhesus monkeys (*Macaca mulatta*). *Proc. Natl. Acad. Sci. U.S.A.* 89, 12137–12139
40 Jensen, K. (2010) Punishment and spite: the dark side of cooperation. *Philos. Trans. R. Soc. Lond. B* 365, 2635–2650
41 Bshary, R. *et al.* (2008) Pairs of cooperating cleaner fish provide better service quality than singletons. *Nature* 455, 964–966
42 Bshary, A. and Bshary, R. (2010) Interactions between sabre-tooth blennies and their reef-fish victims: effects of enforced repeated game structure and local abundance on victim aggression. *Ethology* 116, 681–690
43 Baldassarri, D. and Grossman, G. (2011) Centralized sanctioning and legitimate authority promote cooperation in humans. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11023–11027
44 Kummerli, R. (2011) A test of evolutionary policing theory with data from human societies. *PLoS ONE* 6, 1–6
45 Cant, M.A. (2010) The role of threats in animal cooperation. *Proc. R. Soc. Lond. B* 278, 170–178
46 McNamara, J.M. and Leimar, O. (2010) Variation and the response to variation as a basis for successful cooperation. *Philos. Trans. R. Soc. Lond. B* 365, 2627–2633
47 Oates, J. *et al.* (2010) Roving and service quality in the cleaner wrasse *Labroides bicolor*. *Ethology* 116, 309–315
48 Oates, J. *et al.* (2010) The shadow of the future affects cooperation in a cleaner fish. *Curr. Biol.* 20, R472–R473
49 Pavlov, I.V. (1928) *Lectures on Conditioned Reflexes: the Higher Nervous Activity of Animals*, Lawrence and Wishart
50 Skinner, B.F. (1938) *The Behavior of Organisms*, Appleton-Century Crofts
51 de Quervain, D.J.F. *et al.* (2004) The neural basis of altruistic punishment. *Science* 305, 1254–1258
52 Stevens, J.R. *et al.* (2005) Evolving the psychological mechanisms for cooperation. *Annu. Rev. Ecol. Syst.* 36, 499–518
53 Reiter, J. *et al.* (1978) Northern elephant seal development: the transition from weaning to nutritional independence. *Behav. Ecol. Sociobiol.* 3, 337–367
54 Raihani, N.J. and Bshary, R. (2011) Resolving the iterated prisoner's dilemma: theory and reality. *J. Evol. Biol.* 24, 1628–1639
55 Bull, J.J. and Rice, W.R. (1991) Distinguishing mechanisms for the evolution of cooperation. *J. Theor. Biol.* 149, 63–74
56 Herre, E.A. *et al.* (1999) The evolution of mutualisms: exploring the paths between conflict and cooperation. *Trends Ecol. Evol.* 14, 49–53
57 Noë, R. (2007) Despotic partner choice puts helpers under pressure? *Behav. Process.* 76, 120–125
58 Leimar, O. and Hammerstein, P. (2010) Cooperation for direct fitness benefits. *Philos. Trans. R. Soc. Lond. B* 365, 2619–2626
59 Weyl, E.G. *et al.* (2010) Economic contract theory tests models of mutualism. *Proc. Natl. Acad. Sci. U.S.A.* 107, 15712–15716
60 Pellmyr, O. and Huth, C.J. (1994) Evolutionary stability of mutualism between yuccas and yucca moths. *Nature* 372, 257–260
61 Kiers, E.T. *et al.* (2003) Host sanctions and the legume–rhizobium mutualism. *Nature* 425, 78–81
62 Wong, M.Y.L. *et al.* (2007) The threat of punishment enforces peaceful cooperation and stabilizes queues in a coral-reef fish. *Proc. R. Soc. Lond. B* 274, 1093–1099
63 Noë, R. (2001) Biological markets: partner choice as the driving force behind the evolution of cooperation. In *Economics in Nature. Social Dilemmas, Mate Choice and Biological Markets* (Noë, R. *et al.*, eds), pp. 93–118, Cambridge University Press
64 Johnstone, R.A. and Bshary, R. (2002) From parasitism to mutualism: partner control in asymmetric interactions. *Ecol. Lett.* 5, 634–639
65 Johnstone, R.A. and Bshary, R. (2008) Mutualism, market effects and partner control. *J. Evol. Biol.* 21, 879–888
66 Ferriere, R. *et al.* (2002) Cheating and the evolutionary stability of mutualisms. *Proc. R. Soc. Lond. B* 269, 773–780
67 Foster, K.R. and Wenselaars, T. (2006) A general model for the evolution of mutualisms. *J. Evol. Biol.* 19, 1283–1293
68 Bshary, R. and Schäffer, D. (2002) Choosy reef fish select cleaner fish that provide high-quality service. *Anim. Behav.* 63, 557–564
69 Hamilton, W.D. (1964) The genetical evolution of social behaviour. I. *J. Theor. Biol.* 7, 1–16
70 Hamilton, W.D. (1964) The genetical evolution of social behaviour. II. *J. Theor. Biol.* 7, 17–52
71 Hardin, G. (1968) The tragedy of the commons. *Science* 1, 243–253
72 Gintis, H. *et al.* (2003) Explaining altruistic behavior in humans. *Evol. Hum. Behav.* 24, 153–172
73 Bowles, S. and Gintis, H. (2004) The evolution of strong reciprocity: evolution in heterogeneous populations. *Theor. Pop. Biol.* 65, 17–28
74 McGregor, P.K. (2005) *Animal Communication Networks*, Cambridge University Press
75 Trivers, R. (2004) Mutual benefits at all levels of life. *Science* 304, 964–965
76 Hammerstein, P. and Hagen, E.H. (2006) The second wave of evolutionary economics in biology. *Trends Ecol. Evol.* 20, 604–609
77 Sigmund, K. (2007) Punish or perish? Retaliation and collaboration among humans. *Trends Ecol. Evol.* 22, 593–600
78 Kummerli, R. *et al.* (2010) Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10125–10130
79 Dieckmann, A. (1985) Volunteer's dilemma. *J. Confl. Resol.* 29, 605–610
80 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
81 Boza, G. and Szamado, S. (2010) Beneficial laggards: multilevel selection, cooperative polymorphism and division of labour in threshold public goods games. *BMC Evol. Biol.* 10, 336–348
82 Archetti, M. and Scheuring, I. (2011) Coexistence of cooperation and defection in public goods games. *Evolution* 65, 1140–1148
83 Seymour, B. *et al.* (2007) The neurobiology of punishment. *Nat. Rev. Neurosci.* 8, 300–311
84 Gaston, A.J. (1978) The evolution of group-territorial behavior and cooperative breeding. *Am. Nat.* 112, 1091–1100
85 Mulder, R.A. and Langmore, N.E. (1993) Dominant males punish helpers for temporary defection in superb fairy-wrens. *Anim. Behav.* 45, 830–833
86 Reeve, H.K. (1992) Queen activation of lazy workers in colonies of the eusocial naked mole-rat. *Nature* 358, 147–149
87 Balshine-Earn, S. *et al.* (1998) Paying to stay or paying to breed? Field evidence for direct benefits of helping behavior in a cooperatively breeding fish. *Behav. Ecol.* 9, 432–438
88 Creel, S. (2001) Social dominance and stress hormones. *Trends Ecol. Evol.* 16, 491–497
89 Carlson, A.A. *et al.* (2006) Cortisol levels are positively associated with pup-feeding rates in male meerkats. *Proc. R. Soc. Lond. B* 273, 571–577
90 Young, A.J. *et al.* (2006) Stress and the suppression of subordinate reproduction in cooperatively breeding meerkats. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12005–12010
91 Bshary, R. *et al.* (2011) Short-term variation in the level of cooperation in cleaner wrasse *Labroides dimidiatus:* implications for the role of potential stressors. *Ethology* 117, 246–253