Biology & Philosophy, 27: 833-850

The evolution of punishment

Hisashi Nakao[1] • Edouard Machery[2]

[1] Department of Systems and Social Informatics, Graduate School of Information Science, Nagoya University, Furocho, Chikusa-ku, Nagoya 464-8601, Japan

[2] Corresponding author: Department of History and Philosophy of Science, University of Pittsburgh, 1017 CL, Pittsburgh, PA 15260, USA e-mail: machery@pitt.edu

Abstract: Many researchers have assumed that punishment evolved as a behavior-modification strategy, i.e. that it evolved because of the benefits resulting from the punishees modifying their behavior. In this article, however, we describe two alternative mechanisms for the evolution of punishment: punishment as a loss-cutting strategy (punishers avoid further exploitation by punishees) and punishment as a cost-imposing strategy (punishers impair the violator's capacity to harm the punisher or its genetic relatives). Through reviewing many examples of punishment in a wide range of taxa, we show that punishment is common among plant and animal species and that the two mechanisms we describe have often been important
for the evolution of punishment.

Keywords: Punishment, Cooperation, Moral norms, Evolution of morality

What is the evolutionary function of punishment? While evolutionary theorists often hold that punishment was selected for because it is a means for motivating the punishee to modify its behavior in a way that is beneficial to the punisher, in this article we examine two alternative mechanisms for the evolution of punishment. Looking at a range of taxa, we argue, first, that punishment is common among plant and animal species, a phenomenon neglected by philosophers of biology, and, second, that in many taxa the modification of the punishee's behavior was not crucial to the evolution of punishment[1].

[1] We do not claim that the evolution of punishment was never influenced by the benefits resulting from the

Other researchers have defended the hypothesis that the evolution of punishment did not necessarily depend on benefits derived from the punishee modifying its behavior, and we owe much to their work (e.g., Cant and Johnstone 2006; Bergmu¨ller et al. 2007; West et al. 2007b, 2011; West and Gardner 2010). The main goal of this article is to defend this hypothesis further by clarifying the two possible hypotheses about the evolution of punishment and by looking at many examples in a wide range of taxa.

Here is how we will proceed. Section 1 clarifies the key notions used in this article, including the notion of punishment. Section 2 reviews the two main competing hypotheses about the evolution of punishment. Section 3 shows that in plants and insects the evolution of punishment did not depend on the punishee modifying its behavior in a way that is beneficial to the punisher. Because vertebrates, including mammals, sometimes modify their behavior in response to punishment, punishment in these taxa may be more in line with the idea that punishment is a means for shaping behavior. However, focusing on non-human vertebrates in Sect. 4, we argue that for most species this is probably not the case. We do not think that punishment evolved as a behavioral modification strategy in the human lineage either, but we need more space than we have to make the case for human beings convincingly, and we will focus on that issue in a follow-up article.

Conceptual clarifications

The use of "punishment" across disciplines, and sometimes within a particular discipline, to refer to phenomena that, while superficially similar, are really importantly different, can create much confusion when one theorizes about the evolution of punishment. In this first section, we propose to distinguish these phenomena and to characterize explicitly what we will mean by "punishment" in the remainder of this article. Once this is done, we will be able to identify several forms of punishment.

A behavioral characterization of punishment

Punishment is an action[2] that harms another organism ("the recipient"). Because

punishee modifying its behavior. Rather, looking at many diverse taxa, we argue that this was often not the case.

[2] In this article, the concept of action should be understood very broadly so as to apply to plants, bacteria,

punishment will be discussed in an evolutionary context, harm will be measured in terms of fitness: An organism is harmed (or, as we will occasionally say, a cost is imposed on it) if and only if its absolute fitness is diminished. The reduction of the fitness of the harmed individual is to be thought of as being short-term (in contrast to a reduction of the life-long absolute fitness of the harmed organism). So, in the long-term some organisms may be better off by being punished[3].

Punishment differs from other forms of harming in that it is conditional: Punishment is elicited by some specific harmful action (or trait) performed by the punished organism[4]. We will call the behavior or the trait that triggers either punishment or an inclination to punish a "violation." Some other forms of harming are not conditional: They are not elicited by a particular behavior of the harmed animal[5]. Typically, punishment is elicited by a failure to cooperate, i.e., to behave in a way that is costly to the actor, but beneficial to the recipient. To anticipate some of the examples we will discuss in more detail below, punishment occurs in the context of mutualistic interactions among plants, insects, and fish, in which actors and recipients exchange cooperative behaviors. For instance, in a mutualism between bacteria and soybeans, punishment occurs in a context where fixing nitrogen is costly to the bacteria, but beneficial to the soybeans, while photosynthesizing nutrients is costly to the soybeans, but beneficial to the bacteria.

In addition to comparing punishment with harming in general, it is also useful to clarify the relations between the concepts of spite and punishment (on spite, see Hamilton 1970; Trivers 1985; Foster et al. 2001; Gardner and West 2004a, 2006;

---

etc.

[3] Our reviewers objected that the notion of short-term fitness made no sense. However, first, whether or not this notion makes sense depends on how fitness is conceptualized. In particular, this notion makes perfect sense if fitness is conceptualized as a propensity (roughly, a probabilistic disposition to survive and reproduce), as many philosophers of biology have proposed (e.g., Beatty and Mills 1979) since the strength of this disposition can vary across time. Second, in using this notion we are simply following some evolutionary biologists' discussion of punishment and cooperation (e.g., Bergmüller et al. 2007, 64; West et al. 2007a, R666; West et al. 2011, 235). Third, if the reader's views about fitness are incompatible with the notion of short-term fitness, the notion of a decrease in short-term fitness may be replaced with the notion of immediate payoff loss.

[4] Harassment (e.g., Gilby 2006 for harassment in chimpanzees) is thus not a form of punishment since the harassed individual (e.g., those chimpanzees that possess some food) did not cause any harm.

[5] This definition has the disadvantage to apply to some forms of conditional harming that do not count as punishment, such as defense against predators. We note that it is difficult to define punishment so as to include all actions that intuitive count as punishment and to exclude traits such as defense against predators.

Lehmann et al. 2006; Gardner et al. 2007; West and Gardner 2010). Here, we follow West and Gardner (e.g., West et al. 2007b) in treating as spiteful any trait that reduces the life-long absolute fitness of the actor and of its recipient. It stands in contrast with altruism, selfishness, and mutual benefit (see Table 1). A trait is altruistic if and only if it reduces the life-long absolute fitness of the actor while increasing the life-long absolute fitness of its recipient. A trait is egoistic or selfish if and only if it increases the life-long absolute fitness of the actor while reducing the life-long absolute fitness of its recipient. A trait is mutually beneficial if and only if it increases the life-long absolute fitness of the actor and of its recipient.

Punishment may be spiteful or selfish[6]. It is spiteful when it reduces the long-term absolute fitness of the actor, while being selfish when it increases it. Importantly, what matters for assessing whether punishment is spiteful or selfish is the long-term fitness of the actor, not its shot-term fitness. Even if punishment reduces the fitness of the punishing actor in the short term, it is selfish if its cost for the actor leads to a long-term increase in fitness (e.g., because the punished organism starts acting altruistically toward the punishing actor in response to punishment).

Table 1 Four Types of Traits (see West et al. 2007b)

		Actor	
		Increases its fitness	Decreases its fitness
Recipient	Increases its fitness	Mutual Benefit	Altruism
	Decreases its fitness	Selfishness	Spite

[6] West et al. (2007b, 422) write that "[p]unishment can be selfish or altruistic (like cooperation) or even spiteful, and so without detailed analysis of particular situations, the word 'punishment' should not be given a prefix such as 'altruistic'." Punishment is altruistic when it reduces the long-term fitness of the actor, but increases the long-term fitness of other organisms, e.g., apparented organisms.

We have proposed a purely behavioral, or functional, characterization of punishment: Punishment consists in imposing a cost on another organism as a response to a harmful action. So characterized, the capacity to punish does not require any grasp of norms: It does not require that the punisher be aware that some actions are forbidden or required. It does not even require that the punisher expects the punishee to behave in a particular way or to have or fail to have some specific traits. Indeed, as we have characterized it, punishment does not even require a mind: As some examples discussed in Sect. 3 will show, plants can punish. A behavioral characterization of punishment is required to examine punishment across many different taxa. Building cognitive or psychological concepts into the notion of punishment would lead, by definitional fiat, to the conclusion that only a few mammalian species, perhaps only human beings, punish. We would then be unable to look at many diverse taxa (for a similar defense of behavioral definitions in an evolutionary context, see Thornhill and Thornhill (1989) on rape; Caro and Hauser (1992) on teaching; Wrangham and Peterson (1996) on violence and warfare; for discussion, see Mitchell (2003)).

The characterization of punishment offered here is, to a large extent, in line with influential formulations, especially in evolutionary contexts (e.g., Boyd and Richerson 1992; Clutton-Brock and Parker 1995; Boyd et al. 2003; Cant and Johnstone 2006; Jensen 2010). In their influential article, Clutton-Brock and Parker (1995) define punishment as follows:

> [I]ndividuals (or groups) commonly respond to actions likely to lower their fitness with behaviour that reduces the fitness of the instigator and discourages or prevents him or her from repeating the initial action. We refer to behavioural tactics of this kind as punishment, without implying either a conscious decision or a moral sense on the part of the punisher (209, emphasis added).

Like ours, this definition is strictly behavioral, and it insists on the conditional nature of punishment. However, in contrast to ours, it builds into the notion of punishment the idea that punishment results in a decrease in the frequency of the violation, and it highlights the idea that this decrease results from the punishee modifying its

behavior in response to punishment (punishment "discourages").

Jensen et al. (2007b, 13046) define punishment "in the biological sense […] as a strategy that decreases the occurrence of a behavior," adding that "it is typically selfish in that it provides a future benefit for the individual such as the reduction of harmful behavior received from others." This definition, which just like ours is purely behavioral, differs from ours in tying punishment to a decrease in the frequency of the violation. Jensen (2010, 2637) also defines punishment as "the costly imposition of costs on another individual that result in delayed benefits for the punisher." Just like us, Jensen does not make cognitive capacities a requisite for punishment, but by definition he turns punishment into a selfish action, while our definition leaves this an empirical question.

Finally, Raihani et al. (2012, 288) state that 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." This definition highlights the conditional nature of punishment, but, by stipulation, turns punishment into a costly trait that influences the behavior of other organisms.

Kinds of punishment

To get a sense of the diversity of the phenomenon under consideration here, it is useful to taxonomize the different forms of punishment.

It is common to distinguish second-party from third-party punishment. The former takes place when the violation harms the punisher. To use a toy example of second-party punishment, if a customer tries to leave a restaurant without tipping the waiter, the waiter may lash out at the customer. Third-party punishment takes place when the punisher is not harmed by the violation committed by the punishee. Witnesses of crimes or injustices sometimes get angry, are motivated to punish the criminal, and follow through, even when they have not been in any way harmed: This is an instance of third-party punishment.

Punishment can be open or covert. Punishment is open when the punishee is aware of being imposed a cost; it is covert when this is not the case. Our prototypical representation of punishment involves an individual hurting another one, such as a parent slapping a child, a fight between two individuals, or an executioner executing a prisoner:

These are all instances of open punishment. But punishment need not be open as happens when an organism is not aware of the cost imposed on it as a response to a particular action or to a particular trait.

Punishment can, but need not, be costly: The short-term absolute fitness of the organism can be lowered when it punishes. And, naturally, different kinds of punishment are more or less costly.

Finally, punishment can take place in pairs or in groups. This distinction is important when one is modeling the evolution of punishment. When punishment takes place in a group, benefits the group members, and is costly to the punisher, punishment is a public good and gives rise to the classic free-rider problem: It is in the interest of each member not to punish and to let other group members punish (Oliver 1980; Boyd and Richerson 1992; Henrich and Boyd 2001; Gardner and West 2004b). This issue does not occur in pairs.

Two evolutionary hypotheses about punishment

The main goal of this article is to attempt to characterize the evolutionary pressures that have influenced the evolution of punishment in many different taxa. In this section, we put forward two hypotheses about these pressures, and we examine their initial plausibility.

Punishment as a behavior-modification strategy

One of the most noticeable effects of punishment among humans is its capacity to influence the behavior of the punishee: When she is aware of being punished, and when she understands why she has been punished, the punishee often changes her behavior, and may thus either stop the violation or may be less likely to repeat it. While recidivism is not uncommon, among male convicts who have had less than two convictions and who have been sent to jail for the first time, a majority (53.6%) are not rearrested in a three-year period after their release from jail[7], and parents punish or used to punish their children by physical means (e.g., spanking or whipping) in the hope that children would not commit again the actions that got them punished. Humans are not the only animals

[7] Data from the Bureau of Justice Statistics (BJS) obtained in December 2011 (http://bjs.ojp.usdoj.gov).

whose behavior changes in response to punishment. For instance, punishment is regularly used to modify the behavior of pets.

In light of punishment in humans, it is natural to consider the following hypothesis about the evolution of punishment[8]. Punishment evolved because, as a response to punishment, the punishee modifies its behavior or trait, and the violation stops. If punishment is less costly to the actor than the harm imposed on her by the violation, then, plausibly, selection would favor a disposition to punish violations (but see above on the problems raised by punishment in groups). We will say that, if this hypothesis is correct, punishment is a behavior-modification strategy: Punishment evolved because it modifies harmful behavior. On this view, one may say, punishment is for educating.

The hypothesis that punishment is a behavior-modification strategy is common. First, although some have acknowledged that punishment could evolve for other reasons, they have argued that in most cases punishment did evolve because of its capacity to modify the behavior of the punishees, probably on the grounds that "there is abundant empirical evidence that organisms do respond to punishment" (Boyd and Richerson 1992 177; see also Gardner and West (2004b; West et al. 2011). Clutton-Brock and Parker (1995) write:

> …in the longer run, punishment is a form of selfish behaviour which benefits the punisher because it reduces the probability that the victim will repeat a damaging action or will refuse to perform a beneficial one. In most cases, punishers benefit because victims learn to avoid repeating damaging behaviour, although, in extreme cases, punishment may be beneficial because victims are eliminated or forced to emigrate (209, emphasis added).

While Clutton-Brock and Parker concede that in a few cases, such as eviction, punishment can be beneficial independently of whether the punishee modifies its behavior, they emphasize the role of punishment as a behavior-modification strategy. Clutton-Brock and Parker's view has been extremely influential. For instance, citing Clutton-Brock and

[8] Gardner and West (2004b, 754) are explicit that humans provide the prototype for this approach: "We discuss our model in terms of humans because this is where much of the recent theoretical literature has been focused. However, the implications are general and could be applied to a variety of organisms."

Parker (1995), West et al. (2011, 239) write that "the punished individuals change their behaviour in response to punishment and are more likely to cooperate with the punisher in future interactions" (see also Cant and Johnstone 2006).

Second, many models of the evolution of cooperation (viz. of traits, including behaviors, that are costly to the actor, but beneficial to others) in the human lineage assume that punishment leads the punishees to modify their behavior and forces them to cooperate (Boyd and Richerson 1992; Henrich and Boyd 2001; Boyd et al. 2003, 2010; Gardner and West 2004b). What is more, the evolution of punishment itself is connected to the benefits derived from its capacity to induce cooperation in that these benefits offset the costs inherent in many forms of punishment. Gardner and West (2004b, 753) write that "punishing behavior is often costly for the punisher, and so it is not immediately clear how costly punishment could evolve. […] the crucial factor is a positive correlation between the punishment strategy of an individual and the cooperation it receives. This could arise in several ways, such as when facultative adjustment of behavior leads individuals to cooperate more when interacting with individuals who are more likely to punish."

Punishment without behavioral modification

Although the hypothesis that punishment is a behavior-modification strategy is widespread, there are some alternative hypotheses for the evolution of punishment. What they have in common is the idea that the evolution of punishment does not require benefits produced by behavioral modification. They differ from one another in that they postulate different evolutionary mechanisms to explain how punishment could evolve in the absence of behavioral modification.

The first kind of evolutionary mechanism applies to cases where an actor provides a benefit to a recipient, whose production is costly (in the short term) to the actor itself. Punishment happens when the actor stops providing this benefit. The recipient loses a benefit—thus, a cost is imposed on it—while the actor stops incurring a (self-imposed) cost—it thus gains from punishment. Prima facie, the selection of this form of punishment seems straightforward: When the actor does not benefit from providing a benefit to the recipient, a non-punisher would keep incurring a cost, while a punisher would stop incuring this cost. The latter would thus be better off than the former. We will call this

mechanism a loss-cutting strategy.

Bergmüller et al. (2007, 64) have called a similar mechanism "negative pseudoreciprocity," which they describe as follows: "[A]n individual [punisher] terminates an interaction to avoid immediate fitness losses if the partner [violator] does not invest (or overexploits), thereby stabilising future benefits.'' Cant and Johnstone (2006, 1383) call the same mechanism "self-serving punishment," which "arises where players are allowed to terminate the interaction in response to defection."[9]

We now turn to a second kind of evolutionary mechanism: Imposing a cost on the violator impairs its capacity to harm the punisher or its genetic relatives. In this kind of circumstances, plausibly, punishment can evolve if the cost of punishment for the actor is smaller than the gain derived from reducing the harm caused by the violator (weighted by the coefficient of relatedness when the recipients of the reduced harm are genetic relatives of the punisher). A punisher would be better off than a non-punisher because by incurring a short-term cost it would either get a long-term direct benefit (if the punishee is less able to harm the punisher itself or its descendants) or an indirect benefit (if the punishee is less able to harm genetically related individuals)[10]. In this case, we will say that punishment is a cost-imposing strategy. In a sense, we may say, punishment is for harming.

There are different versions of the evolutionary mechanism under consideration. First, the punishee may be killed or so badly harmed that it is unable to harm the punisher anymore. Second, the punishee's fitness may be reduced, and the punishee may thus leave fewer descendants (which would also be violators). As a result, there will be fewer descendants able to harm the descendants of the punisher (if the punisher and punishee have a similar life expectancy) or to harm the punisher itself (if the punishee has a shorter life cycle than the punisher). In all these cases, punishment provides a (long-term) direct benefit to the punisher. Because the shot-term cost of punishment is thus offset, punishment can evolve, and it is then selfish (see the quotation of West et al. 2007b earlier). Third, instead of having direct benefits, punishment can have indirect benefits: It can impair the capacity of the punishee to harm genetic relatives of the punisher or it can

[9] This very simple idea also can be extended to interactions in a larger group. See Hirshleifer and Rasmusen (1989) for a formal treatment.

[10] See West et al. (2007b) for a useful clarification of the distinction between direct and indirect benefits.

reduce the number of descendants left by the punishee, which will result in less harm imposed on genetic relatives of the punisher. In this case, punishment is spiteful.

The hypothesis that the evolution of punishment does not require any behavioral modification has been endorsed by some evolutionary biologists concerned with the initial evolution of punishment. For instance, Cant and Johnstone (2006, 1383) argue that "[a] strategy of punishment can invade a population of nonpunishers if it pays regardless of the response of the other player," adding that, after punishment has evolved, it begins to function as a threat and behavior-modification strategy. Nowak and Highfield (2010, 223) also argue that punishment in the human lineage "is not…a mechanism for the evolution of cooperation" but "…fits neatly into the framework of reciprocity," i.e., punishment evolved as a strategy "where I suffer a loss but, as a result of his, you suffer a bigger one." While Gardner and West focus mostly on the view that punishment is a behavior-modification strategy, they also consider the third version of the cost-imposing strategy (West et al. 2007b, 422). What these citations have in common is the idea that punishment can evolve without behavioral modification. They highlight different alternative mechanisms for the evolution of punishment: While Cant and Johnstone refer to self-serving punishment or negative pseudoreciprocity—i.e., punishment as a loss-cutting strategy—Nowak and Highfield as well as West et al. highlight selfish or spiteful punishment—i.e., punishment as a cost-imposing strategy. These ideas should naturally be elaborated formally to show that they are really plausible (taking into account features such as population structures, noise in whom gets punished, etc.). However, our goal in the remainder of this article will be different: We will present a large amount of empirical evidence suggesting that the evolution of punishment in numerous lineages did not depend on the punishee modifying its behavior, but rather on the kind of mechanisms just described.

While the role of behavioral modification distinguishes the scenarios just considered and the scenario examined at the beginning of Sect. 2, these scenarios have all in common the idea that the evolution of punishment requires some correlation between punishment and some form of (direct or indirect) benefit. They disagree about the source of the benefit that is correlated with punishment: a change in the behavior of the punishee in response to punishment, the punisher ceasing to harm itself, and a reduction in the punishee's capacity to harm.

Initial plausibility of the behavior-modification hypotheses

The hypothesis that punishment is a behavior-modification strategy may seem initially plausible in light of the universality of operant conditioning in the animal kingdom (Raihani et al. 2012). Operant conditioning (aka stimulus–response learning) is the following phenomenon: When an organism (e.g., a rat or a sea slug) behaves in a particular way and when its behavior results in some temporally contiguous harm, it is less likely to behave this way again (e.g., Thorndike 1901; Skinner 1953). For instance, a rat in a Skinner box that presses a lever and gets electrically shocked is less likely to press the lever again. Operant conditioning is widespread among animals: It is found in insects such as flies, mollusks such as sea slugs, birds such as pigeons, and mammals such as rats, primates, and humans. This suggests that it evolved early in the history of life. Now, one could speculate that the behavioral flexibility imparted by operant conditioning was instrumental in the evolution of punishment: Organisms that happened to harm those that harmed them (i.e., violators) would have a fitness advantage over those that don't since, by operant conditioning, violators would be less likely to harm again. While the speculation linking punishment to operant conditioning may see initially plausible, it suffers from a severe limitation. For operant conditioning to be the basis of the evolution of punishment, punishment would have to be temporally contiguous to the violation; otherwise, punishment would not negatively condition the organism—viz. decrease the probability that it would commit the violation again. The problem is that, as we shall see in Sect. 3, in many taxa punishment often is not temporally contiguous to the violation.

Where does this discussion leave us? It would seem that the plausibility argument just considered fails to favor clearly the hypothesis that punishment is a behavior-modification strategy. In the remainder of this paper, we thus move beyond this type of plausibility arguments to examine punishment in a range of taxa[11].

---

[11] Our reviewers noted that we did not discuss the possibility that the function of punishment is deterrence. The reason is that, for most of the cases considered in this article, deterrence is more likely to be a by-product of the evolution of punishment than its evolutionary function. That is, in these cases, after punishment evolved for a reason other than deterrence, punishees then evolved to modulate their behavior as a function of the capacity and likelihood of the victim of a violation to punish (Cant and Johnstone 2006).

Punishment in plants and insects

Although some evolutionary biologists have looked at punishment in many different lineages other than humans (e.g., Bshary and Grutter 2005 and Wong et al. 2007 on fishes; Kiers et al. 2003 on soybeans; Young et al. 2006 on meerkats; see also West et al. 2007b for a review), recent discussions of the evolution of punishment in philosophy of biology and other disciplines have tended to focus on the human lineage and on primates (e.g., Boyd et al. 2003, 2010; Henrich et al. 2006; Sripada 2005; Rohwer 2006; Jensen et al. 2007a, b; Mathew and Boyd 2011). This narrow focus has unfortunately led philosophers of biology as well as some scientists to overlook the fact that punishment is found in many taxa, including among plants, insects, fish, and mammals. In this section, we examine punishment in plants and insects, and we show that its evolution does not necessarily depend on the punishee modifying its behavior.

Punishment is common among plants and insects

Surprisingly perhaps, punishment is known to occur among various species of plants. A striking instance of punishment is found in the mutualistic interaction between rhizobia (bacteria) and soybeans (Kiers et al. 2003). Rhizobia live in nodules attached to the root of soybeans and fix nitrogen within them. While fixing nitrogen is costly to the rhizobia, fixed nitrogen contributes to the growth of soybeans, which cannot fix nitrogen by themselves. In return for the fixed nitrogen, soybeans provide the rhizobia with nutrients produced by photosynthesis, which is costly to the soybeans. This mutualistic interaction is not perfect since some lineages of rhizobia do not fix nitrogen. They benefit from a trait that is costly to soybeans (providing nutrients) without benefiting these in return, and they impose a cost on soybeans. When soybeans recognize cheating rhizobia[12], they reduce the oxygen permeability of the root nodules, and this reduction reduces the amount of

---

This scenario is likely because, for deterrence to be the function of punishment, the punishee would have had to be sufficiently behaviorally flexible to modulate their behavior in response to punishment. This is unlikely to be the case for most of the cases we consider in this article. (Recall that we do not discuss punishment among human beings).

[12] Soybeans do not always recognize the lineages of cheating rhibozia (Kiers et al. 2003, 79).

nutrients available to cheating rhizobia. Stopping the provision of nutrients to cheating rhizobia benefits soybeans because they thereby stop an activity that is costly to themselves—providing nutrients—and punishing soybeans are better off than non-punishing soybeans that are exploited by cheating rhizobia for a longer time. This is an instance of the loss-cutting strategy described in Sect. 2. In some respects, this is also a cost-imposing strategy. Stopping the provision of nutrients is costly to the rhizobia: Their number decreases significantly when soybeans reduce the oxygen permeability of the root nodules. Soybeans benefit because fewer cheating rhizobia are able to harm them or their relatives.

Cordia nodosa (Boraginaceae) makes hollow swellings at branch internodes (called ‘‘domatia’’), which are good lodging for some species of ants (e.g., Azteca ants), and sometimes provide them with food (Edwards et al. 2006). While the production of domatia is costly, it is repaid by the ants patrolling and protecting Cordia nodosa from herbivores. However, some species of ants live in the domatia without patrolling. These ants benefit from a trait that is costly to the plant (producing domatia) without benefiting it in return, and, thus, they impose a cost on the plant. Cordia nodosa punishes such cheater species by stopping the growth of the domatia when its leaves, being not protected, end up being heavily damaged. Punishment is beneficial to Cordia nodosa because it stops producing domatia. This is another instance of a loss-cutting strategy. In some respects, stopping the growth of domatia is also a cost-imposing strategy. This punishment is costly to the cheating ants that lose housing and sometimes food. It is efficient: Almost 90% of the species found on Cordia nodosa are protecting species.

There are several other cases of punishment in mutualistic interactions between plants and insects, such as fig trees and fig wasps (Jandér and Herre 2010) and yucca and yucca moths (Pellmyr and Huth 1994). In the mutualistic interactions studied by Jandér and Herre, each species of fig trees has one or a few species of fig wasps as specific partners: These wasps carry the pollen of their partner fig trees (which is costly to the wasps and beneficial to the fig trees), while their eggs are laid within the figs, where their larvae can live (which is costly to the fig trees and beneficial to the wasps). Some wasps cheat by laying eggs within the figs without carrying the pollen. When this happens, the fig trees punish the cheating wasps by aborting the figs in which the cheating wasps lay eggs. This punishment leads to the death of the larvae within the aborted fig, which

seriously reduces the fitness of the cheating wasps, as is shown by the fact that the proportion of cheaters among wasps increases as the rate of abortion decreases (Jandér and Herre 2010, 1485, figure 3). Punishment is efficient: Even in Ficus popenoei, the species of fig tree with the smallest rate of abortion, one finds only very few cheating wasps specialized to this fig tree (around 5%). Punishment is not very costly to the fig trees, and it is beneficial: They stop being exploited by cheater wasps. This case of punishment seems to be both a loss-cutting and a cost-imposing strategy.

Just like many species of plants, some species of insects also punish cheaters (e.g., West-Eberhard 1986; Monnin and Ratnieks 2001; Ratnieks and Wenseleers 2008). For instance, female paper wasps (Polistes dominula) have a rigid social hierarchy, in which an alpha female lays eggs. They can identify other individuals' ranks by means of the shape of their spots on the clypeus: The "brokenness" of these spots is positively correlated with body size and social dominance, and these spots signal wasps' rank—they are "badges of status." If these badges are to be useful as rank indicators, they have to be honest: If it were possible to fake them, these badges would be unreliable signals, and wasps would not use them to signal their rank. In many species, this kind of signal has some physiological costs, as is the case for peacocks' feathers (e.g., Zahavi, 1975; Grafen, 1990). However, because the spots on paper wasps' clypeus do not impose significant costs on their holders, paper wasps have another mechanism for preventing cheating. When they find cheaters with broken spots that are disproportionate to their body sizes (e.g., small individuals having more broken spots), they seriously punish them: Low-ranking individuals experimentally colored with more broken spots receive approximately six times more aggression from higher-ranking individuals than low-ranking individuals that, as a control, are colored in a way that does not change their visual appearances (Strassmann 2004; Tibbetts and Dale 2004). This type of aggression increases the risk of injury and is costly to the cheaters. For our purposes, what is important in this example is that cheaters do not modify the shape of their spots in response to punishment since the violating traits are not behavioral, but morphological. So the evolution of punishment in this case does not seem to depend on any modification of the punishes' behavior: Because of such punishment, cheaters will just be selected against, and status signals will remain useful. This is an instance of a cost-imposing strategy.

Punishment and non-responsive behavior

As we have seen in Sect. 1, since Clutton-Brock and Parker (1995), it has often been assumed that punishment evolved because the cost of punishment is compensated with benefits gained from the punishees modifying their behavior (e.g., by starting to cooperate). However, the examples of punishment in plants and insects we have just reviewed show that punishment can evolve even in the absence of behavioral modification.

What these examples have in common is that the violators do not stop imposing costs in response to punishment: Rhizobia do not start fixing nitrogen when the soybeans stop providing them with nutrients; cheating ant species do not start patrolling when Cordia nodosa stop the growth of domatia; fig wasps do not start carrying pollen when the figs are aborted; and paper wasps do not modify their morphological traits. For this reason, punishment cannot have evolved in these taxa because the punishees modify their violating behavior or trait in response to punishment. In these taxa, at least, punishment cannot be a behavior-modification strategy: In soybeans, Cordia nodosa, figs, and some wasps, punishment is not for educating.

In some of these cases, punishment is naturally viewed as a loss-cutting strategy. Consider for instance the case of rhizobia and soybeans. As we have seen, cheating rhizobia, for instance, do not provide benefits to soybeans, and punishment consists in soybeans terminating their costly interaction with them. In other cases, it is naturally viewed as a cost-imposing strategy. Dominant wasps reduce the fitness of cheating wasps, and thus the number of cheating wasps' offspring. Finally, many cases can be seen as instances of both strategies.

Why are rhizobia, ants, and fig wasps non-responsive to punishment? It is naturally not the case that these organisms are entirely inflexible: Just like any organism, their traits or behaviors change as a function of the environment they are in, and we have seen in Sect. 2 that operant conditioning is widespread in the animal kingdom. On the other hand, the flexibility of these organisms is rather limited: They only respond to a limited range of stimuli, and they change their behavior in stereotypical ways. Furthermore, as is the case for paper wasps, some violating traits are largely determined genetically and cannot be modified in response to punishment. It may be that being

responsive to punishment requires a greater flexibility than the one found in these organisms.

Punishment in vertebrates

Punishment has evolved in relatively inflexible, non-responsive organisms possibly because by punishing the punisher stops harming itself or because the punishee (or its relatives) is less able to harm the punisher (or its relatives). On the other hand, the behavioral repertoire of contemporary vertebrates, including mammals, is flexible, and they do respond to punishment (see examples below). It is thus reasonable to ask whether the evolution of increased behavioral flexibility did influence the evolution of punishment: Once more flexible behavior evolved, did punishment evolve differently? It may be that in vertebrates and even more clearly in mammals various features of punishment are best explained if punishment's capacity to modify the behavior of punishees gave rise to new selective pressures. Although this conjecture looks plausible and may explain the evolution of punishment in some taxa, we argue in this section that often the evolution of punishment in vertebrates and mammals probably was not influenced by punishees' capacity to modify their behavior in response to punishment. We first review data about punishment in a range of taxa before discussing their implications for the issue discussed in this article.

Punishment in fish

Punishment is known to occur in several species of fish. In a few species, but not in all, punishment results in a modification of the punishee's behavior, and may be a behavior-modification strategy. We illustrate the two cases, looking first at a species where punishment is not connected to a change in behavior.

Social rank in Paragobiodon xanthosomus (Gobiidae), a coral-reef fish, is stabilized by punishment (Wong et al. 2007). Gobiidae have a social hierarchy based on body size: The largest individual is dominant, the next one is its immediate subordinate, and so on. This rank is relatively stable because the growth of subordinates is regulated and does not exceed a size threshold. If the size of a fish exceeds this threshold, the dominant individual chases and evicts it from the group. This eviction seriously reduces the fitness of the violator given the ecology of these coral-reef fish: It is unlikely that the

evicted individual will be able to move to another coral, and Paragobiodon xanthosomus cannot breed outside corals. Furthermore, although punishing dominants may incur some costs by chasing and evicting the subordinates, these costs are smaller than those they would incur if subordinates deprived them of their corals.

Punishment also plays an important role among cleaner and client fish and among male–female pairs of cleaner fish (Bshary and Grutter 2005; Mills and Cóte 2010; Raihani et al. 2012). The cleaner wrasse, Labroides dimidiatus, eats ectoparasites in the mouth of the client fish, which is beneficial both to the cleaner and the client. However, the wrasse seems to prefer the mucus on their clients' skin to ectoparasites (Bshary and Grutter 2005), but the wrasse is punished for eating its client's mucus: When this happens, the client typically looks for another cleaner fish, but it sometimes also chases the cheater cleaner. Bshary and Grutter (2005) have experimentally shown that cleaner wrasses change their behaviors, and limit themselves to eating ectoparasites after having been chased[13]. Thus, punishment by client fish maintains the mutualism between clients and cleaners. While punishment by client fish may be a behavior-modification strategy, this may not be the case. Because non-punishers will be exploited to a greater extent than punishers, punishers are fitter than non-punishers. In this sense, punishment by clients is similar to the loss-cutting strategies we have examined in the previous section.

Punishment in (non-human) mammals

Whereas mammals and, especially primates, often respond to punishment by modifying their behavior (see the studies discussed below), we argue that the selective forces acting on punishment in these flexible animals probably did not depend on behavioral modification.

Meerkats' cooperative breeding provides a clear-cut example of punishment without behavioral modification. Dominant meerkats exclusively reproduce in the group, a monopoly maintained by the eviction of pregnant subordinates. Subordinates suffer serious physiological stress from this eviction, resulting in a change in hormonal levels and in a higher abortion rate among pregnant subordinates (Young et al. 2006). Thus, eviction is seriously harmful for the subordinates, and it constitutes a kind of punishment,

[13] The same is true of the cooperation between male and female cleaners that work in pairs.

given the notion of punishment put forward in Sect. 1. Importantly for our purposes, this type of punishment is not for educating: The punishees (evicted pregnant subordinate female meerkats) do not modify their behavior in response to punishment; rather, punishment imposes significant costs on the punishees, impairing their capacity to harm (i.e., by producing offspring) the punishers. Pregnancy-related punishment among female meerkats is naturally characterized as a cost-imposing strategy.

Turning now to primates, while rhesus monkeys on Cayo Santiago often produce a call when they find food, they sometimes try to monopolize it. When other rhesus monkeys detect this behavior, they often attack the cheating individual, depriving it from the food (Hauser 1992). Hauser suggests that food sharing is maintained by such punishment (Hauser 1992, 12139). This instance of punishment may look like a good example of a behavior-modification strategy since punishment seems to promote cooperative food sharing. However, first, it is unclear whether rhesus monkeys are more likely to call after having been punished (Raihani et al. 2012). Furthermore, the crucial benefits of punishment need not be derived from the behavioral modification it may induce: Punishers get food as immediate benefits, whether or not those individuals who have tried to monopolize it will produce a call in the future (Stevens et al. 2005, 506–507). The cost of punishment is also usually low since typically only higher-ranked rhesus monkeys attack: When the other rhesus monkeys have a lower rank than the food discoverer, they typically do not attack and simply sit near the discoverers (except when they manage to recruit coalitional support); by contrast, if the other rhesus monkeys have a higher rank than the food discoverer, they attack the food discoverer.

More interestingly, Jensen et al.'s (2007b) experiment suggests that chimpanzees do not always modify their behaviors in response to punishment. In Study 2, a chimpanzee was presented with some food on a tray; another chimpanzee in another cage had the opportunity to drag the tray to itself, and thus to steal the food presented to the first chimpanzee, which they tended to do (74 %of the trials). The first chimpanzee had however been taught that it could make the tray collapse, which would prevent the violator from eating the stolen food. Chimpanzees were more likely to make the tray collapse compared to a condition in which a human experimenter took the food away from the first chimpanzee to give it to the second chimpanzee. Importantly, even though the first chimpanzee punished the second, punishment did not lead to a reduction in theft:

"There was a trend toward an increase in theft, suggesting that retribution did not have an effect on noncooperative behavior" (Jensen et al. 2007b, 13048). So, punishment took place without behavioral modification.

We are aware of the risks of over-interpreting Jensen et al.'s experimental result: In particular, the non-responsiveness of the chimpanzees to punishment may be due to some peculiarities of the experimental situation, and may not be found in the wild. Still, this experiment suggests that chimpanzees do not always respond to punishment, and that (at least non-costly) punishment in chimpanzees is not always conditional on the punishee's responsiveness (apparently, chimpanzees continued punishing despite the inefficiency of punishment). These two features are not predicted by the hypothesis that in the primate lineage punishment was under selection due to the benefits derived from the punishees modifying their behavior or traits. In addition, Raihani et al. (2012) have recently noted that punishment does not seem to modify chimpanzees' behavior in natural environments.

Punishment without behavioral modification

As some of the examples discussed in Sect. 4 suggest, fish and mammals, including primates, sometimes respond to punishment by modifying their behavior: Punished cleaners stop eating mucus, and punished rhesus monkeys may be more likely to call when they find food. So, in some species, punishment may be a behavior-modification strategy. However, a few considerations suggest that punishment among vertebrates was probably not (or at least only rarely) influenced by selective pressures created by the capacity of the punishees to modify their behavior in response to punishment.

First, in the cases considered in this section that involves behavioral modification, punishment would still be beneficial to the punisher even if the punishee did not modify its behavior in response to punishment. Punished cleaner wrasses do not impose any more cost on their clients (by eating their mucus), and punishing rhesus monkeys get immediate benefits whether or not punishment increases rhesus monkeys' likelihood to call in the future.

Second, in the other cases presented in this section, the benefits of punishment for the punisher do not depend on any behavioral modification. Punished coral-reef fish cannot reproduce while punishing coral-reef fish retain their social rank and their capacity

to breed; physiological changes caused by exclusion reduce punished female meerkats' reproduction rate, and their offspring do not compete with the offspring of the punishers; and chimpanzees do not modify their behavior even after having been punished.

Moreover, some of the examples considered in this section are naturally viewed as cost-imposing strategies. For instance, although it is somewhat costly to the punishing dominant coral-reef fish to chase and evict the growing subordinates, punishers are better off than non-punishers who will lose their corals.

Why are there only few cases where punishment can be viewed as a behavior-modification strategy among vertebrates? This may be due to the nature of some cognitive capacities such as long-term memory and temporal discounting (for a similar argument about cooperation, see Stevens and Hauser 2004; Stevens et al. 2005). Temporal discounting consists in "devaluing the future rewards" and it "often results in a preference for smaller, immediate rewards over larger, delayed rewards" (Stevens et al. 2005, 509). Because of higher temporal discounting in nonhuman animals than humans, the former may discount the delayed benefits behavioral modification would bring, and they may not punish cheaters unless they can get immediate benefits. Moreover, if non-human animals' long-term memory is worse than humans, punishees may forget the cost of being punished and may not change their behavior in future interactions.

Upshot

Punishment is common among vertebrates, including fish and mammals. Vertebrates sometimes modify their behavior in response to punishment, but typically punishment would be beneficial even if this were not the case. Furthermore, in many cases, punishment happens whether or not the punishee changes its behavior, while in other cases punishment is beneficial because it imposes costs on the punishee or because it prevents the punishee to impose further costs on the punisher. This suggests that among vertebrates too punishment may typically not be for educating.

Conclusion

The goal of this article was to clarify the notion of punishment and to assess an influential hypothesis about the evolution of punishment—viz. that punishment is for educating. The examples of punishment in plants and insects suggest on the contrary that in many plant

and insect species the evolution of punishment probably had little to do with influencing the behavior of the punishee; rather, punishment seems either for harming or for cutting one's loss. While non-human vertebrates are often responsive to punishment, and adapt their behavior accordingly, in many taxa, punishment was probably not influenced by selective pressures created by vertebrates' responsiveness to punishment.

References

Beatty J, Mills S (1979) The propensity interpretation of fitness. Philos Sci 46:263–288

Bergmüller R, Johnstone RA, Russell AF, Bshary R (2007) Integrating cooperative breeding into theoretical concepts of cooperation. Behav Process 76:61–72

Boyd R, Richerson P (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethol Sociobiol 13:171–195

Boyd R, Gintis H, Bowles S, Richerson P (2003) The evolution of altruistic punishment. Proc Natl Acad Sci 100(6):3531–3535

Boyd R, Gintis H, Bowles S (2010) Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science 328:617–620

Bshary R, Grutter AS (2005) Punishment and partner switching cause cooperative behaviour in a cleaning mutualism. Biol Lett 1:396–399

Cant MA, Johnstone RA (2006) Self-serving punishment and the evolution of cooperation. J Evol Biol 19(5):1383–1385

Caro TM, Hauser MD (1992) Is there teaching in nonhuman animals? Q Rev Biol 67(2):151–174

Clutton-Brock TH, Parker GA (1995) Punishment in animal societies. Nature 373:209–216

Edwards DP, Hassall M, Sutherland WJ, Yu DW (2006) Selection for protection in an ant-plant mutualism: host sanctions, host modularity, and the principal-agent game. Proc R Soc B 273:595–602

Foster KR, Wenseleers T, Ratnieks FLW (2001) Spite: hamilton's unproven theory. Ann Zool Fennici 38:229–238

Gardner A, West SA (2004a) Spite and the scale of competition. J Evol Biol 17:1195–1203

Gardner A, West SA (2004b) Cooperation and punishment, especially in humans. Am

Nat 164:753–764

Gardner A, West SA (2006) Spite. Curr Biol 16:R662–R664

Gardner A, Hardy ICW, Taylor PD, West SA (2007) Spiteful soldiers and sex ratio conflict in polyembryonic parasitoid wasps. Am Nat 169:519–533

Gilby I (2006) Meat sharing among the Gombe chimpanzees: harassment and reciprocal exchange. Anim Behav 71:953–963

Grafen A (1990) Biological signals as handicaps. J Theor Biol 144(4):517–546

Hamilton WD (1970) Selfish and spiteful behaviour in an evolutionary model. Nature 228:1218–1220

Hauser M (1992) Costs of deception: cheaters are punished in rhesus monkeys (Macaca mulatta). Proc Natl Acad Sci 89:12137–12139

Henrich J, Boyd R (2001) Why people punish defectors: conformist transmission stabilizes costly enforcement of norms in cooperative dilemmas. J Theor Biol 208:79–89

Henrich J, McElreath R, Barr A, Ensimger J, Barrett C, Bolyanatz A, Cardenas JC, Gurven M, Gwako E, Henrich N, Lesorogol C, Marlowe F, Tracer D, Ziker J (2006) Costly punishment across human societies. Science 312:1767–1770

Hirshleifer D, Rasmusen E (1989) Cooperation in a repeated prisoners' dilemma with ostracism. J Econ Behav Organ 12:87–106

Jandér KC, Herre EA (2010) Host sanctions and pollinator cheating in the fig tree-fig wasp mutualism. Proc R Soc B 277:1481–1488

Jensen K (2010) Punishment and spite, the dark side of cooperation. Philos Trans R Soc B 365:2635–2650

Jensen K, Call J, Tomasello M (2007a) Chimpanzees are rational maximizers in an ultimatum game. Science 318:107–109

Jensen K, Call J, Tomasello M (2007b) Chimpanzees are vengeful but not spiteful. Proc Natl Acad Sci 104(32):13046–13050

Kiers ET, Rousseau RA, West SA, Denison RF (2003) Host sanctions and the legume-rhizobium mutualism. Nature 425:78–81

Lehmann L, Bargum K, Reuter M (2006) An evolutionary analysis of the relationship between spite and altruism. J Evol Biol 19:1507–1516

Mathew S, Boyd R (2011) Punishment sustains large-scale cooperation in prestate

warfare. Proc Natl Acad Sci 108:11375–11380

Mills SC, Cóte IM (2010) Crime and punishment in a roaming cleanerfish. Proc R Soc B 277:3617–3622

Mitchell S (2003) Biological complexity and integrative pluralism. Cambridge University Press, New York

Monnin T, Ratnieks FLW (2001) Policing in queenless ponerine ants. Behav Ecol Sociobiol 50:97–108

Nowak M, Highfield R (2010) SuperCooperators: Altruism, evolution, and why we need each other to succeed. Free Press, New York

Oliver P (1980) Rewards and punishments as selective incentives for collective action: theoretical investigations. Am J Sociol 85:1356–1375

Pellmyr O, Huth CJ (1994) Evolutionary stability of mutualism between yuccas and yucca moths. Nature 372:257–260

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

Ratnieks FLW, Wenseleers T (2008) Altruism in insect societies and beyond: voluntary or enforced? Trends Ecol Evol 23:45–52

Rohwer Y (2006) Hierarchy maintenance, coalition formation, and the origins of altruistic punishment. Philos Sci 74(5):802–812

Skinner BF (1953) Science and human behavior. Macmillan, New York

Sripada CS (2005) Punishment and the strategic structure of moral systems. Biol Philos 20:767–789

Stevens JR, Hauser MD (2004) Why be nice? Psychological constraints on the evolution of cooperation. Trends Cogn Sci 8(2):60–65

Stevens JR, Cushman FA, Hauser MD (2005) Evolving the psychological mechanisms for cooperation. Ann Rev Ecol Evol Syst 36:499–518

Strassmann JE (2004) Rank crime and punishment. Nature 432:160–162

Thorndike EL (1901) Animal intelligence: an experimental study of the associative processes in animals. Psychol Rev Monogr Suppl 2:1–109

Thornhill R, Thornhill N (1989) An evolutionary analysis of psychological pain following rape: I. The effects of victim's age and marital status. In: Betzig L (ed) Human nature: a critical reader. Oxford University Press, New York, pp 225–238

Tibbetts EA, Dale J (2004) A socially enforced signal of quality in a paper wasp. Nature 432:218–222

Trivers R (1985) Social evolution. Benjamin-Cummings, New York

West SA, Gardner A (2010) Altruism, spite and greenbeards. Science 327:1341–1344

West SA, Griffin AS, Gardner A (2007a) Evolutionary explanations of cooperation. Curr Biol 17:R661–R672

West SA, Griffin AS, Gardner A (2007b) Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J Evol Biol 20:415–432

West SA, El Mouden C, Gardner A (2011) 16 common misconceptions about the evolution of cooperation in humans. Evol Hum Behav 32:231–262

West-Eberhard MJ (1986) Dominance relations in Polistes Canadensis (L.), a tropical social wasp. Monitore Zool Italiano (Nuova Serie) 20:263–281

Wong MYL, Buston PM, Munday PL, Jones GP (2007) The threat of punishment enforces peaceful cooperation and stabilizes queues in a coral-reef fish. Proc R Soc B 274:1093–1099

Wrangham RW, Peterson D (1996) Demonic males: Apes and the origins of human violence. Houghton Mifflin, Boston

Young AJ, Carlson AA, Monfort SL, Russell AF, Bennett NC, Clutton-Brock T (2006) Stress and the suppression of subordinate reproduction in cooperatively breeding meerkats. Proc Natl Acad Sci 103(32):12005–12010

Zahavi A (1975) Mate selection: a selection for a handicap. J Theor Biol 53(1):205–214