Love or Fear: Can Punishment Promote Cooperation?

SEBESTIAN KROUPA

Cooperation is a paradox: Why should one perform a costly behavior only to increase the fitness of another? Human societies, in which individuals cooperate with genetically unrelated individuals on a considerably larger scale than most mammals do, are especially puzzling in this regard. Recently, the threat of punishment has been given substantial attention as one of the mechanisms that could help sustain human cooperation in such situations. Nevertheless, using punishment to explain cooperation only leads to further questions: Why spend precious resources to penalize free-riders, especially if others can avoid this investment and cheaters can punish you back? Here, it is argued that current evidence supports punishment as an efficient means for the maintenance of cooperation, and that the gravity of proposed limitations of punishment for maintaining cooperation may have been overestimated in previous studies due to the features of experimental design. Most notably, the importance of factors as characteristic of human societies as reputation and language has been greatly neglected. Ironically, it was largely the combination of the two that enabled humans to shape costly punishment into numerous low-cost and less detrimental strategies that clearly can promote human cooperation.

“Upon this a question arises: whether it be better to be loved than feared or feared than loved? It may be answered that one should wish to be both, but, because it is difficult to unite them in one person, […] either must be dispensed with. […] Love is preserved by the link of obligation which, owing to the baseness of men, is broken at every opportunity for their advantage; but fear preserves you by a dread of punishment which never fails.” Niccolo Machiavelli, The Prince, ch. XVII. p 57–58.

In his last presidential address to the Royal Society in 2005, Robert May claimed that “the most important unanswered question in evolutionary biology, and more generally in the social sciences, is how cooperative behaviour evolved and can be maintained.” Humans are particularly puzzling in this regard, as they are known to cooperate with genetically unrelated individuals on a substantially larger scale than do most mammal species.2–4 Recently, the notion of costly punishment has been hotly discussed as one of the factors that could boost human cooperation.2,5–8

Following Clutton-Brock and Parker (1995), punishment is a "retaliatory infliction of fitness reduction … 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."9,209 This definition is now widely accepted.6,10,11 Punishment is therefore a temporarily spiteful strategy, which nonetheless yields delayed benefits because the recipient’s behavior becomes more cooperative in subsequent interactions (Fig. 1).

To illustrate this strategy, one of the best pieces of evidence of punishment in nonhuman species comes from the mutualistic interaction between the blue-streak cleaner wrasse (Labroides dimidiatus) and its fish clients. Cleaners are expected to remove ectoparasites and dead tissue, but occasionally cheat by feeding on mucus on their client’s skin, thus harming the client. In a series of experiments, it has been observed that if clients have been cheated, they often aggressively chase the cleaner.12 Although it is instantly costly to the client, this behavior brings clear future benefits, since cleaners thus punished are less likely to cheat in subsequent interactions.

PUNISHMENT AMONG HUMANS

Some of the strongest evidence that humans are motivated to punish cheaters and that punishment promotes cooperation comes from diverse economic games. First, in the ultimatum game, a sum of money is allotted to two anonymous players and one of them proposes how to split the stake between the two. The
second player can either accept the offer, in which case the players keep the money, or reject it, in which case both players receive nothing. To avoid reciprocation, the game is played only once. The second player, if acting according to economic rationality, should accept any nonzero offer, as the chimpanzee (Pan troglodytes) typically does in this game. When the game is played by humans, however, most offers below 20%–30% are rejected.14,15 This trend has been observed across a range of cultures from all six continents and thus seems to be a universal human trait (Fig. 2). In light of this evidence, one could argue that people are willing to punish unfair behavior even if the penalty is costly in terms of resources.

The third-party punishment game presents a similar scenario. Three anonymous players receive a sum of money, half of which is allocated to the third player. Player one then proposes how to split the remaining half between the first two players. While the second player is relegated to the role of a mere spectator, the third player can invest a part of his stake to decrease player one’s payoff, should he consider his offer to player two unfair. Although a self-interested or neutral player three should never pay to penalize player one, cross-cultural studies indicate that, as in the ultimatum game, offers below 20%–30% are commonly punished (Fig. 3).15,16 Consequently, even as uninterested third parties, humans appear to be willing to inflict costly punishment.

While results from the ultimatum and third-party punishment games indicate that people are motivated to punish even if such penalties are costly, evidence from the public goods game demonstrates that the punishment of selfish individuals promotes and stabilizes cooperation at higher levels than if no penalties are available.2,5,17,18 In the public goods game, players can either cooperate by contributing to the common pool, which is multiplied and evenly distributed among all members at each turn, or defect, thereby refusing to share their endowment. Self-interested individuals typically realize that by withholding their tokens, they can achieve higher payoffs and that, in consequence, average contributions quickly decrease close to zero. Nevertheless, if costly punishment of other players is allowed — that is, if individuals can pay a cost to reduce the payoff to others — free-riders are, heavily penalized by the group and average contributions approach the maximum value (Fig. 4).

Further findings highlight the strength of the motivation to punish, since humans impose costly penalties on cheaters regardless of whether they engage in repeated or one-shot anonymous encounters, although the latter promise no possibility of subsequent cooperative interactions.18 Moreover, humans are eager to pay even for only symbolic punishment, or rather, disapproval, which does not directly harm the free-riding individual.19 On the other hand, penalties are, to some degree, still subject to economic rationality, since the willingness to inflict punishment decreases as its cost grows.8,20,21 It has been documented that the degree of costly punishment positively co-varies with the extent of cooperation across cultures,15 supporting the notion that penalties are important in the maintenance of cooperation. It might seem that the positive effect of punishment on sustaining cooperation can be hardly doubted.

The Hazardous Nature of Costly Punishment

“In the state of nature there often wants power to back and support the sentence when right, ... resistance many times makes the punishment dangerous, and frequently destructive, to those who attempt it.”

John Locke, Second Treatise of Civil Government91:66.

Although ample evidence supports the role of punishment in the maintenance of cooperation, this strategy
has been widely criticized as too costly and destructive. Punishment is effective in promoting cooperation, since the threat of harm makes virtually everyone contribute highly. However, cooperation enforced by punishment is largely inefficient because the costs of punishment may exceed the benefits of cooperation. Once costly punishment is extensively used, average group payoffs may decrease considerably, since both the punishing and the punished lose resources. In consequence, net payoffs in groups with punishment are often lower than payoffs in groups without the option to punish (Fig. 5).

Moreover, punishers are subject to exploitation by second-order free-riders who do cooperate, but fail to pay to penalize cheaters, and therefore achieve higher payoffs than those who punish. In this respect, consider the following question: Why sacrifice for the public good and punish, especially if the act is costly? Obviously, some individuals, even if they are otherwise cooperative, might be tempted to avoid this investment.

Furthermore, although initial models and studies investigating the effects of costly punishment on cooperation had typically assumed that penalties are targeted exclusively at cheaters, this need not be the case. Studies have now shown that punishment can be directed not only at free-riders, but also at cooperators or punishers. This so-called antisocial punishment interferes with the beneficial influence of the punishment of cheaters, hindering both the effectiveness and efficiency of cooperation, which breaks down and yields even lower benefits. As punishers of cheaters themselves become the target of punishment, they obviously are less likely to punish free-riders in return.

This phenomenon is especially prevalent when an additional round of punishment is added to the game by allowing counter-punishment. As punishers of cheaters themselves become the target of punishment, they obviously are less likely to punish free-riders in return.

Experimental evidence clearly shows that antisocial punishment presents a real threat to cooperation. Typically, 20% or more of all punitive acts are directed at high contributors, while about 25% of punishments targeted at cheaters are retaliated. Antisocial punishment is common across a wide range of cultures, correlating negatively with the strength of rule of law and civic norms of cooperation. Therefore, using costly punishment to explain the motivation behind cooperation leads to further questions: Why spend precious resources to penalize others, especially if second-order free-riders can avoid this investment?
and anti-social punishers can punish you back?

Punishment in the Lab: Time Heals Everything

The crucial issue of present research entails its relevance to real-life situations. Because of a mismatch between experimental conditions and the situations in which humans typically use punishment in everyday circumstances, the efficiency of punishment could be inadvertently underestimated. In many cases, evidence that punishment undermines the efficiency of cooperation has come from studies in which the public goods game ran for ten to twenty rounds.\textsuperscript{21,22,25,30,34} However, if the game is allowed to run over a higher number of periods, although average earnings are initially higher in the penalty-free condition, punishment becomes less detrimental over time, eventually yielding cooperation that is both highly effective and efficient (Fig. 6).\textsuperscript{8,18,35,36} This occurs as a result of a steep decline in the frequency of punishment over successive rounds (Fig. 7).

Punishment still remains available as an option when it is necessary to discipline occasional defectors, but there is no longer any need for its extensive use. Since the resources available are not depleted on extensive penalties and the degree of contributions constantly remains high, overall payoffs increase. However, even if rarely applied, punishment still holds the key for cooperation: as soon as it is removed, cooperation deteriorates.\textsuperscript{2,5} In the long run, the beneficial influence of punishment on the effectiveness of cooperation persists, but its negative effect on efficiency wanes away.

Why does the frequency of punishment drop over time? The likely explanation is rather simple: People gradually learn the rules of the game or, in other words, punishment requires several rounds to modify the behavior of its recipients, establish as a shared expectation, and reach equilibrium.\textsuperscript{8,37} Once this has occurred, participants typically have observed or directly learned that free-riding will not be tolerated by the group. They also have become aware that there is no benefit in cheating. The actual enactment of punishment thus gradually declines as the threat of its infliction is established. Experimental evidence shows that repeated interactions are likely important in this process, allowing the formation of expected future behavior, as well as credibility between players.\textsuperscript{8}

The initial rounds of the game could thus be regarded as artificial, since the collective strategy has not yet been agreed on or enforced through punishment. This also means that if punishment is truly effective in sustaining cooperation, it should quickly reach equilibrium, where penalties are only seldom used. Therefore, the absence of extensive, costly punishment in a given population could suggest an especially strong local effect; that is, that most cheaters have already modified their behavior and do not need to be penalized again. This could be the case of real-world populations where, since common...
expectations have typically been established and cheaters know what to expect, punishment should be carried out rarely and thus bear minimal overall costs. The availability of the option to punish, however, is essential if this state is to be maintained. To illustrate this point, although the vast majority of taxpayers are never fined for tax evasion, the existence of this measure clearly discourages people from cheating. As experimental studies demonstrate, if the threat of punishment is diminished or removed, tax evasion increases.38–40

**No Choice But to Punish**

Experiments investigating punishment commonly take the form of a public goods game in which one can pay to impose costs on cheaters. However, in the absence of an alternative strategy, the frequency of costly punishment could be accidentally overrepresented, since participants have no other option to discipline cheaters except through costly punishment.41,42 In other words, since no intermediate form of punishment is available, players can choose only between two radical options: use a highly costly strategy that might induce overly strong damage or abstain from using it. While the former destroys a large portion of the resources available and negatively affects the efficiency of cooperation, the latter cannot sustain cooperation at all.

This point bears special relevance to modeling studies, which frequently use evidently inaccurate payoff matrices. In this respect, it is essential to realize that punishment satisfies its objective when it reduces the cheater’s payoff to just below the population average: at this precise point, cooperation becomes more profitable than free-riding.43 Therefore, any penalties stronger than this are unnecessary and merely damage the public good without soliciting higher contributions, while any investments lower than this yield no benefits at all. Such misrepresentation of the relationship between the costs and the benefits of punishment could lead to an underestimation of its efficiency.

Similarly, in real-world conditions, punishment is not limited to the two extremes and intermediate forms are available. Moreover, situations outside of the lab offer additional alternatives that could help control the hazardous nature of punishment, such as communication and freedom of choice, whereby participants can agree on their decisions, expectations and punitive rules. Accordingly, as the current body of evidence suggests, higher payoffs are achieved if punishment is combined with communication than if the opportunity for dialogue is absent.44,45

Studies allowing endogenous group formation and rule selection, where players can decide whether to join a group in which costly punishment is allowed or banished, and discuss how and against whom to use penalties, illustrate how costly punishment is handled if players are not as constrained by the experimental design but, instead, given the freedom to choose between strategies and partners.33,44,46 This direction of research solved the essential question of whether people would deliberately adopt costly punishment.

The answer is affirmative, but with a curious twist. Initially, participants are usually reluctant to allow costly punishment of others or join a group with punitive rules. However, the experience of free-riding gradually convinces them that such measures are necessary if cooperation is to be sustained.33,44,47 At the end of the game, virtually all groups would adopt costly punishment or all players would migrate to a group in which the option of punishment is available. Moreover, these groups would achieve the highest payoffs and lowest levels of defection.47

No group ever voted to authorize penalties on high contributors, thus instantly banning antisocial punishment.33 When participants were allowed voluntarily to associate with players of their preference, and thus determine their own group composition every few turns, cooperators gathered together and free-riders were forced to act as cooperators if they wished to be involved at all.46 Nonetheless, the option of penalizing free-riders was equally vital, as only the combination of regrouping and costly punishment yielded the strongest cooperation. Interestingly, merely the effect of choice could boost cooperation. In spite of there being essentially no difference between endogenously and exogenously implemented rules, it has been shown that the former yields higher payoffs.48

The key advantage of the design discussed could be that since participants can openly agree on their shared strategy or join a group where rules have been firmly established, the aforementioned initial artificial phase of the experiment, in which no collective strategy exists, is skipped or
considerably reduced and equilibrium is reached instantaneously. In other words, participants do not need to form common expectations of cooperation and punishment through trial and error, as in the classic form of a public goods game, but instead arrive at them through the faster and cheaper means of communication and deliberation. The results of studies allowing for endogenous decision making could be thus interpreted as supporting the previously discussed notion of equilibrium. Considering that communication is normally available in real-life situations, it would be desirable to take this factor into account when investigating the effect of punishment on cooperation among humans.

How Relevant Is Second-Order Free-Riding?

As stated, nonpunishing cooperators avoid the cost of punishment and thus achieve higher average payoffs than do punishing cooperators. Therefore, the question is why sacrifice for the general good and punish. Obviously, some individuals, termed second-order free-riders, might be tempted to avoid this investment and thus achieve higher payoffs. The essence of this problem is based on the assumption that the benefits of punishment are shared equally among all members of a given population.\textsuperscript{49,50} However, this is most likely incorrect. In terms of reputation,\textsuperscript{26,51} for example, a disproportionate share of the benefits could erase or surpass the difference in payoffs between punishing and nonpunishing cooperators and thus solve the issue.

Although precise evidence for such additional benefits is absent, current studies demonstrate only negligible occurrence of second-order free-riding in humans. This behavior provoked very little anger among participants who were largely unwilling to penalize such behavior.\textsuperscript{32,52} Moreover, an observational field study in the Ju/'hoansi Bushmen showed that second-order free-riding never became a subject of conflict.\textsuperscript{53} Briefly, the issue appears to bear little or no relevance to real-world circumstances.

The question is, therefore, how humans evolved to solve this problem. On one hand, its significance may have been exaggerated by the use of supposedly inappropriate payoff matrices that assume equal distribution of the benefits of punishment and therefore disregard, for example, the effect of reputation. However, coordination and institutions could also hold the key, since the whole community may be expected to participate and share the costs or, perhaps, appoint and eventually grant professional status to specific punishers.

AntiSocial Punishment: Hiding Behind Anonymity?

When interpreting evidence from studies investigating the effects of antisocial punishment, it is impossible not to consider the role of anonymity. The vast majority of experiments on punishment are conducted with subjects under conditions of strict anonymity. Players typically communicate by computer and are informed of the contributions of others to them on-screen, with their identities subsequently reshuffled.\textsuperscript{2,5} Therefore, it is truly impossible for players to establish any permanent links with other players and their behavior, which prevents any influence on reputation.

The impact of anonymity on the extent of antisocial punishment remains unclear and controversial. Critics of costly punishment argue that anonymity protects punishers from receiving additional vengeful penalties from the cheaters they had previously punished, so that the rules of the experimental design conceal the full effect of antisocial punishment.\textsuperscript{23,30,41} However, anonymity might well contribute to exaggeration of the incidence of antisocial punishment. It has been long known that the perception of anonymity combined with feelings of reduced individual responsibility can produce high rates of misconduct, including aggressive or punitive actions.\textsuperscript{54,55} As indicated, experiments investigating antisocial punishment typically take place in analogous conditions.

Further, the use of technology can encourage antisocial behavior that individuals otherwise probably would not engage in, such as online trolling.\textsuperscript{56,57} On that note, perceptual differences when interacting with a human or a computer partner in economic games have been reported, with the former provoking a stronger neurocognitive response.\textsuperscript{58} Consequently, it seems that anonymity could protect antisocial punishers rather than punishers of cheaters.

To resolve this debate, it is essential to do further experiments, ideally in conditions where participants cannot hide behind anonymity. Although studies have been carried out under such circumstances,\textsuperscript{46,59} none have looked directly at the interaction between lack of anonymity and antisocial punishment. Yet, the introduction of identity, reputation, and personal contact seems unlikely to favor antisocial behavior.

The Power of Forgiveness

Retaliatory aggression is frequent among most social animals,\textsuperscript{9} so why, in light of Mahatma Gandhi’s famous quote, is not the whole world blind? In this regard, reconciliation, a common and simple strategy that could bear particular relevance to vengeful antisocial punishment among humans, has been overlooked. This behavior, which has been documented among social primates and social canines, typically is employed to minimize the risk of conflict escalation.\textsuperscript{9,60–62} Anthropological field evidence demonstrates that reconciliation has a role in human populations as well. Among the Tsimané people of lowland Bolivia, individuals prefer to solve their conflicts through reconciliation rather than revenge.\textsuperscript{63} A reconciliation body, such as a council of elders, which presents another alternative, is used by the Pathan Hill tribes in Afghanistan.\textsuperscript{64} Analogously, the most respected figures and leaders among the Ju’hoansi Bushmen often refrain from taking part in the most severe punitive conflicts so that they can mediate reconciliation between the adversaries and prevent an escalation of aggression.\textsuperscript{53} It seems clear that reconciliation is rather difficult without some sort of
communication. One could argue that this behavior is greatly facilitated by the presence of language among humans. Nevertheless, since verbal or, generally speaking, personal interaction is typically unavailable in experiments investigating both costly and antisocial punishment, reconciliation is inaccessible. Consequently, punitive conflicts might needlessly escalate. Taking this into account, conducting a study in which both the opportunity of antisocial punishment and communication are enabled could simulate real-world conditions more accurately.

Punishment Outside of the Lab: The Many Faces of Costly Penalties

During the last fifteen years, intensive investigations of punishment, largely based on economic dilemma games, have clearly demonstrated that humans have a strong, universal motivation to punish cheaters in order to push them to cooperate. Nonetheless, as noted, the central issue with experiments investigating the relationship between punishment and cooperation in humans is their rigid nature. Punishment is often unrealistically considered as unconditional, uncoordinated, and directly provoked by defection. Since factors as characteristic of human societies as reputation and communication are commonly eliminated, the gravity of the problems typically associated with the costs of punishment is, perhaps, exaggerated.

In spite of all that, there remain some limitations on the power of punishment in the maintenance of cooperation. Clearly, punishment is a potentially dangerous tool, breeding aggression and cycles of retaliation that could lead to disarray, violence, or social breakdown. However, drawing on indirect reciprocity, coordination, and verbal reproach, humans have devised mechanisms to tame and control the hazardous nature of punishment and forged it into less destructive forms. As a result, punishment in real-world conditions is typically coordinated, associated with reputation, mediated by language and, briefly, less detrimental.

Indirect Reciprocity: When Punishment and Reputation Pull Together

“A good name is to be chosen rather than great riches, and favour is better than silver or gold.”

Proverbs 22:1

The main currency that indirect reciprocity brings into play is represented by reputation. One does not need to consult academic literature to figure out that reputation greatly matters in our social interactions and is relevant to punishment and cooperation: We despise cheaters and are more likely to punish in the presence of larger audiences. Similarly, if reputation is allowed to form during experiments, general group-beneficial behavior, such as cooperation, is positively affected. It appears also that the neurocognitive system involved in processing cooperation and punishment overlaps with the one assessing reputation. Merely observing the face of a partner with the reputation of a cooperator activates the same reward-related areas in the brain as punishment and cooperation do.

Looking at the neurocognitive basis of evaluating the faces of cheaters could be also insightful.

There seem to be at least two distinct ways that punishment and reputation can interact to promote cooperation. First, punishment can directly increase the reputation of punishers, making them more likely to receive cooperation in subsequent interactions as well as turning them into more desirable mates or allies.

Second, indirect reciprocity allows an individual to select cooperative partners and shun uncooperative ones based on their reputations; it thus provides an additional incentive to cooperate and reduces the need to use costly penalties. Since cheaters suffer from low regard, cooperators might selectively withhold their cooperation from them and direct it toward individuals who are reputable cooperators instead. In consequence, as harmful interactions with free-riders are avoided, punishment need not be performed that frequently, which minimizes its overall costs and its negative effect on the efficiency of cooperation. Whereas the first option could be termed “building the reputation of a punisher,” the second one might be called “building the reputation of a cooperator.” Do we, therefore, build the former, the latter, or both?

Building the Reputation of a Punisher

It has been suggested that punishers gain an indirect benefit in terms of reputation, which directly compensates for the costs of punishment and thus could equally alleviate the second-order free-riding problem. In this respect, reputable punishers should be less likely to be cheated or denied cooperation in future interactions and, at the same time, more likely to be regarded as desirable social and mating partners. But do punishers actually accrue benefit from their higher reputation?

The answer to this question is affirmative. In a computer public goods game, where the roles of punisher, free-rider, and nonpunisher were predetermined, players rated the former as more trustworthy, group-focused, and worthy of respect than the latter two. Moreover, in subsequent iterated trust games, where participants were paired with one of the preprogrammed players, punishers enjoyed the highest trust and amassed highest earnings. Nevertheless, participants were sensitive to the fairness of punishment, as only justifiable cases of imposing penalties on free-riders were rewarded with increased reputation. There is evidence of a similar process in a real-world population. Among the Bergdama, a seminomadic people of Namibia, especially costly cases of punishment are usually carried out by younger men who, if successful, benefit from increased prestige.

Nonetheless, punishment can sometimes become too much of a good thing and prove detrimental to reputation. Among the Ju’hoansi Bushmen, individuals who engage in
overly frequent or harsh punishment are regarded as inconsiderate and do not benefit from high reputation. Most esteemed figures or leaders, accordingly, refrain from taking part in severe punitive conflicts and draw their reputation from subsequent mediation rather than punishment. By the same token, when managers in companies are observed punishing their subordinates, their employees sometimes report that they are less trustworthy and respectful.

The relationship between punishment and reputation is therefore far from straightforward. It is clear that to a certain extent it pays off to be a punisher, in terms of both reputation and future interactions. Unjustified, overly harsh, or antisocial punishment, on the other hand, appears to damage a punisher’s reputation. As stated, punishment need not always be the most preferable strategy in terms of reputation. Its appropriateness seems to depend on the context of a situation. The specifics of the link between reputation and punishment should become a priority in future research.

Building the Reputation of a Cooperator

Indirect reciprocity equally allows one to build the reputation of a cooperator and interact with other reputable cooperators while shunning cheaters. By withdrawing cooperation, one can easily terminate or avoid harmful interactions at little cost. Defectors, on the other hand, are “punished” through social rejection or loss of help and mating opportunities. As cheaters are cut off from the cooperative circles, the detrimental effect of free-riding on these circles is mitigated and cooperation can thrive. In addition, as encounters with free-riders are intentionally avoided, punishers need not often inflict costly penalties, which mitigates the negative effect of punishment on the efficiency of cooperation.

An experimental study by Rockenbach and Milinski allowed an interaction between withdrawal of cooperation and costly punishment. As expected, only the combination of both strategies yielded the most effective and efficient cooperation. Since participants were naturally concerned with building the reputation of a cooperator, indirect reciprocity provided an alternative incentive to cooperate, which in turn reduced the need for extensive use of costly penalties in order to enforce cooperation. Availability of the option to punish and discipline occasional defectors, however, was required if this effective and efficient cooperation was to be sustained. In the absence of costly penalties, when cooperation relied solely on reputation, its efficiency and effectiveness declined.

The strategy of withdrawing cooperation can draw on rich anthropological evidence. In the Pathan Hill tribes of Afghanistan, recurrent cheaters suffer from covert social shunning: for example, they are avoided in forming alliances and never invited to participate in the Council of Elders. Similarly, among the Ju'/hoansi Bushmen, stingy or uncooperative individuals are held in low social regard, and thus have fewer marriage or social exchange opportunities. The importance of reputation for finding a mate is demonstrated by a study in the Turkana, a nomadic society of Kenya, where, when asked about their feelings toward a fictional cowardly warrior, unmarried women tended to be more critical than any other respondents.

It seems clear that indirect reciprocity and costly punishment mutually reinforce cooperation. Although punishment can become a secondary measure under the rule of indirect reciprocity, its presence and occasional infliction are required if cooperation is to be sustained at a high level. To conclude and return to the initial question, humans conceivably build both the reputation of a cooperator and a punisher, since both bring benefits. The relative importance of the two should be investigated by future research.

Reputation Is Not Everything

Indirect reciprocity appears to be a very important mechanism in the maintenance of human cooperation. But what if information about reputation is unavailable or unreliable? Presumably, this is not an uncommon scenario in the real world. In this respect, as population size increases, information quality of systems based on reputation rapidly declines, thus making indirect reciprocity a doubtful candidate for sustaining cooperation at large scales. In consequence, indirect reciprocity appears to bear more relevance for cooperation in the circle of close acquaintances for whom reputation is either available or can be easily acquired. To sustain cooperation beyond this circle, however, additional mechanisms, such as coordinated punishment and institutions, might be required.

Coordination: United We Stand, Divided We Fall

As the evidence from studies allowing endogenous decision-making demonstrates, if players are permitted to communicate and determine their own rules, cooperation thrives. In this respect, group control and coordination of penalties bring numerous benefits. In the first place, coordination helps decrease the costs of punishment, which are either distributed among the whole community or managed in the cheapest way possible. Moreover, if a rule is unified and backed by the group’s majority or dominant members, its legitimacy is established, which reduces the likelihood of antisocial punishment. Collective punishment thus provides the necessary control for the potentially detrimental effects of costly penalties, minimizing both their costs and the occurrence of antisocial punishment. This could equally help alleviate the second-order free-riding problem, since everyone may be expected to participate and share the costs or, alternatively, punishers can be appointed by the community. As ethnographical studies indicate, punishment is rarely unconditional and uncoordinated in real-world populations. The Turkana, a nomadic people of Kenya, sustain large-scale cooperation in combat through coordinated costly punishment.
and others. In other words, unre-
nal policing forces, judges, mediators, 
les and penalties in institutions that establish firm 
agement eventually gives rise to 
coordinated punishment, espe-
ally by those harmed by the cheating 
action bears costs in terms of time and 
and may damage social rel-
ations, yet the communal consent and 
ordination ensure legitimacy 
and efficacy.

Control imposed by the commu-
ity seems to be essential for the cor-
rect function of costly punishment. 
In the South African tribes of Berg-
dama and !Kung, as well as among 
igenous Australians, serious 
gressions of rules are broadly 
debated before a final decision is 
ached and suitable punishers nom-
ned. To prevent vengeful 
ishment, individuals responsible 
 for mediation may be appointed or, 
 alternately, the act of punishment 
ay be delegated to the violator’s 
. Coordination seems especially 
relevant to the most costly forms of 
ishment. Among the Ju/hoansi 
shmen, the vast majority of mild 
disapproval is administered by indi-
iduals, but sharp criticism is mostly 
divered by alliances. Similarly, 
afety in numbers is more important 
when punishing particularly corrupt 
behavior that immediately threatens 
community stability.

In brief, anthropological field evi-
dence suggests that human coopera-
tion at larger scales, where 
reputation gradually loses its power, 
is to a great extent enforced by third-
party coordinated punishment, espe-
cially as regards highly costly penal-
ties or particularly dangerous forms 
of cheating. Although coordinated 
ishment still bears its costs, these 
are greatly reduced by the power of 
community. In societies that operate 
on extremely large scales, coordina-
tion of penalties eventually gives rise 
to institutions that establish firm 
control over crime and punishment, 
legitimize rules and penalties in 
codes of laws, and employ profes-
ional policing forces, judges, mediators 
and others. In other words, unre-
stricted costly penalties are explicitly 
prohibited and punishment is cen-
tralized in the hands of the state, 
with cheats or antisocial punishers 
actively prosecuted for their offen-
ces. Therefore, formation of institu-
tions promoting cooperation in 
everyday interactions might have 
been especially crucial for the de-
velopment of large-scale cooperation in 
humans. This possibility is supported 
by the fact that both market integra-
tion and the presence of a world reli-
ion positively co-vary with fairness 
 across different cultures.

The Irresistible Urge to Gossip

As ethnographic evidence indi-
cates, punishment in real-world pop-
ulations typically starts as talk and 
gradually escalates to more severe 
forms: from joking, mocking, and 
mild complaints to sharp criticism 
and eventually violent acts. 
Wiessner’s investigation of punish-
ment among the Ju/hoansi Bushmen 
fittingly demonstrates the impor-
tance of verbal reproach. Out of 171 
cases of punishment recorded, 98% 
were verbal (22% mocking and jok-
ing, 41% mild complaints, 35% sharp 
criticism) and only 2% of punish-
ments were represented by costly 
violent acts. Although it is debata-
able to what extent Wiessner’s obser-
vations can be considered as punish-
ment, she reports that among the 
Ju/hoansi, verbal reproach, which is 
typically triggered by disrup-
tive behavior and idleness, can 
bear substantial costs for both the 
actor and the recipient and, in many 
cases, leads to a change in the recipi-
ent’s behavior — which fits the defi-
nition of punishment.

One specific form of verbal pun-
ishment has received considerably 
more attention than any other: gos-
ip. It is no secret that chitchat is 
central to our nature: it has been 
estimated that gossip could account 
for 65% of all human conversation. 
Through sharing diverse information 
about an absent third party and 
combining reputation and coordina-
tion, gossip enables fast and cheap 
communication of news across the 
whole community. This constant 
flow of reputational information 
allows close surveillance of the 
behavior of others. Because of its 
particularly strong effect on reputa-
tion, gossip can promote cooperation 
in a number of ways.

First, by spreading reputational 
information about cheats who are 
then mocked or shunned, gossip 
mediates the process of building the 
reputation of a cooperator and ena-
bles one to selectively withhold coop-
eration from defectors. Second, by 
increasing reputational incentives, 
gossip provides extra motivation to 
build the reputation of being a coop-
operator and refrain from cheating. 
In this regard, it has been found that in 
the presence of a bystander who, as 
conveyed to participants, was likely 
to gossip about the behavior observed, 
cooperativeness in eco-
omic games greatly increased.

Third, by showing that one is willing 
to spread negative information if 
cheated, gossip not only deters 
others from defecting, but also con-
tributes to formation of the reputa-
tion of a punisher. While gossip 
can be equally abused to spread false 
information, liars and overly nega-
tive or frequent gossipers are likely 
to be distrusted and become targets 
of gossip themselves.

Although verbal punishment is typi-
cally regarded as bearing little or no 
price and thus as not being a strat-
egy associated with punishment, this 
perspective greatly neglects the 
impact of social and psychological 
costs. In this respect, the outcome of 
overheard gossip in terms of dam-
aged social relationships can be 
quite severe. Similarly, among the 
Ju/hoansi Bushmen, sharp criti-
cism typically involved emotional 
pain and even material consequences 
for both the punishing and the pun-
ished individual. Very frequently, 
an open tension between the two 
remained and they would not engage 
in any sort of social exchange until 
the conflict was solved. Yet despite 
the well-known risks, the Ju/hoansi 
are not afraid to verbally express 
their disagreement.

In addition to costs, one more clue 
relates gossip to costly punishment. 
The negative affective response that 
drives punishment seems to be 
equally involved in encouraging gos-
siping behavior, which suggests that
gossip is underpinned by the same motivation that underlies costly punishment in economic games.\textsuperscript{84} Therefore, in light of the evidence presented, it appears that verbal reproach has more in common with punishment than one might initially say. It is frequently triggered by inappropriate behavior, can be rather costly, can elicit a change in behavior, and is probably driven by identical motivations.

Language Is Behind Everything

As argued here, in real-world conditions the motivation to punish can be manifested in many different, often cheaper and safer forms, since costly penalties can wear the mask of reputation, coordination, or verbal reproach. It can be argued that language, in particular, underlies these three factors and might therefore have made a highly significant contribution to the diversification and versatility of punishment in humans.\textsuperscript{4,88} In this regard, the ability to speak greatly facilitates coordination, as it allows groups to define specific rules and laws. Moreover, the ease and speed of humans’ verbal communication permitted a more efficient exchange of social information, for example through gossip, and thus allowed reputation to play a considerably more important role in the life of community. Finally, the immediate cost of both detecting and punishing cheaters is considerably lower if verbal communication and reproach are available. In brief, because of its low price and ability to mobilize collective action, language might be one of the key reasons behind the high success of punitive strategies among humans. It may equally reinforce other elements that promote cooperation among humans, such reputational mechanisms or signaling.\textsuperscript{88}

Since human cooperation occurs in a wide array of ecologies and domains, slightly different measures of control in terms of punitive behaviors might have been favored by selection in different environments.\textsuperscript{77,89} Consequently, in conditions where reputation, coordination, or language provide cheaper options of social control, other, more costly means of punishment may become limited and rarely used.\textsuperscript{59,77} The extraordinary diversity of punitive measures observed in the field testifies to this notion. Not only have real-world human populations presumably reached punishment equilibrium a long time ago, but they also had millennia to devise strategies that reduce the costs of punishment and tame its risks.\textsuperscript{53,77} Among the Ju’hoansi Bushmen, for example, only 2\% of the penalties observed were responses to extremely costly, violent behavior.\textsuperscript{53}

In view of that, it is crucial to bear in mind that punishment need not always be extremely costly to promote cooperation, but it is precisely this eventual option of harsh penalties that backs up the use of milder forms of punishment.\textsuperscript{90}

To illustrate this point, metaphorically speaking, one could imagine costly punishment as a policeman with a gun. Merely his presence is usually sufficient to keep the peace and order, and even when the law is broken, he typically does not need to reach immediately for his gun. Instead, he can start with verbal reproach or threaten to use his weapon before he resorts to highly costly forms of punishment. Even though the gun represents an extreme and a rarely used measure, it grants the policeman additional power and provides him with strong leverage in case it is needed. In its absence, the policeman would lose these advantages, making his job of enforcing rules and maintaining obedience to them considerably more difficult. Therefore, although milder penalties will be used more frequently than costly punishment, the presence of this severe threat is what ultimately sustains their efficiency. As Hobbes wrote in his \textit{Leviathan}, ch. XVII “Covenants, without the
Sword, are but Words, and of no strength to secure man at all”.

CONCLUSION

The last decade of intensive research, predominantly drawing on economic games, has produced strong evidence that people across all cultures readily punish free-riders in order to push them toward cooperation, even if this behavior bears substantial costs. Indeed, the universal presence of a strong motivation to punish cheaters can be regarded as the most important piece of evidence offered by previous empirical efforts. New insights, nevertheless, have gone hand in hand with challenges to the notion that punishment could play a role in the maintenance of human cooperation. It was regarded as too costly and destructive, and as suffering from the problems of second-order free-riding and antisocial punishment.

However, as argued in the first section of this paper, the gravity of the suggested limitations may have been overestimated in early studies as a result of the features of experimental design. Most importantly, the role of factors as characteristic of human societies as reputation and language has been neglected. In the absence of communication, common expectations have to be formed through trial and error, which might undermine the efficiency of punishment, while the problems of second-order free-riding and antisocial punishment may have been exaggerated due to inappropriate payoff matrices and anonymity, which protects deviant individuals. In the second part of the article, we have demonstrated how the combination of communication and indirect reciprocity allowed humans to shape costly punishment into numerous low-cost and less detrimental strategies.

Although the combination of communication and reputation might seem to be able to alleviate most problems associated with costly penalties, strong and precise experimental evidence for some of the claims introduced here is currently missing. Most notably, the specifics of the relationship between punishment and indirect reciprocity — for example, the extent to which punishers benefit from improved reputation — have not been satisfactorily addressed. Second-order free-riding also deserves further consideration: Has the problem been exaggerated by the use of inappropriate payoff matrices disregarding reputation, for example, or did humans evolve a different solution, perhaps in terms of coordination? Moreover, researchers have yet to shed light on the role of anonymity in antisocial punishment, the importance of reconciliation, the construction of common expectations, and the achievement of equilibrium.

In spite of having provided priceless insights, models and experimental games can only take us so far. Without a change in direction, we cannot expect new findings to emerge. Above all, future research should aim to expand on the previous restrictive experimental design and prioritize investigations of punishment in more natural settings, where the influence of and interaction with reputation, communication, and possibly other factors could be studied. Field studies especially have the potential to become a decisive source of data and new knowledge. Only this approach can lead to a more complete picture of costly punishment and its role in the maintenance of large-scale cooperation in humans.

ACKNOWLEDGMENTS

I deeply thank Mike Webster and Sally Street for their guidance and generous comments on earlier drafts of this article. I also thank Andy Gardner and three anonymous reviewers for helpful comments.

REFERENCES

1 May RM. 2006. Address of the President, Lord May of Oxford OM AC FRS, given at the Anniversary Meeting on 30 November 2005: threats to tomorrow’s world. Notes Rec R Soc 60:109–130.
2 Fehr E, Gächter S. 2002. Altruistic punishment in humans. Nature 415:137–146.
3 Fehr E, Fischbacher U. 2003. The nature of human altruism. Nature 425:785–791.
4 Melis AP, Semmann D. 2010. How is human cooperation different? Philos Trans R Soc Lond B Biol Sci 365:2663–2674.
5 Fehr E, Gächter S. 2000. Cooperation and punishment in public goods experiments. Am Econ Rev 90:980–994.
6 Gardner A, West SA. 2004. Cooperation and punishment, especially in humans. Am Nat 164:753–764.
7 Sigmund K. 2007. Punish or perish? Retaliation and collaboration among humans. Trends Ecol Evol 22:593–600.
8 Balliet D, Mulder LB, Van Lange PAM. 2011. Reward, punishment, and cooperation: a meta-analysis. Psychol Bull 137:594–615.
9 Clutton-Brock TH, Parker GA. 1995. Punishment in animal societies. Nature 373:209–216.
10 Raihani NJ, Thornton A, Bshary R. 2012. Punishment and cooperation in nature. Trends Ecol Evol 27:288–295.
11 West SA, Griffin AS, Gardner A. 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J Evol Biol 20:415–432.
12 Bshary R, Gruetter AS. 2002. Asymmetric cheating opportunities and partner control in a cleaner fish mutualism. Anim Behav 63:547–555.
13 Jensen K, Call J, Tomasello M. 2007. Chimpanzees are rational maximizers in an ultimatum game. Science 318:107–109.
14 Henrich J, Boyd R, Bowles S, et al. 2001. In search of Homo economicus: behavioral experiments in 15 small-scale societies. AEA Pap Proc 91:73–78.
15 Henrich J, McElreath R, Barr A, et al. 2006. Costly punishment across human societies. Science 312:1767–1770.
16 Bernhard H, Fischbacher U, Fehr E. 2006. Parochial altruism in humans. Nature 442:912–915.
17 Bowles S, Gintis H. 2004. The evolution of strong reciprocity: cooperation in heterogeneous populations. Theor Popul Biol 65:17–28.
18 Gächter S, Herrmann B. 2009. Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. Philos Trans R Soc Lond B Biol Sci 364:791–806.
19 Carpenter JP, Daniere AG, Takahashi LM. 2004. Cooperation, trust, and social capital in Southeast Asian urban slums. J Econ Behav Organ 55:533–551.
20 Carpenter JP. 2007. The demand for punishment. J Econ Behav Organ 62:522–542.
21 Egas M, Riedl A. 2008. The economics of altruistic punishment and the maintenance of cooperation. Proc R Soc London B, Biol Sci 275:871–878.
22 Dreber A, Rand DG, Fudenberg D, et al. 2008. Winners don’t punish. Nature 452:348–351.
23 Rand DG, Dreber A, Ellingsen T, et al. 2009. Positive interactions promote public cooperation. Science 325:1272–1275.
24 Ohtsuki H, Iwasa Y, Nowak MA. 2009. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature 457:79–82.
25 Herrmann B, Thoni C, Gächter S. 2008. Antisocial punishment across societies. Science 319:1362–1367.
26 Boyd R, Richerson PJ. 2005. Solving the puzzle of human cooperation. In: Levinson SC, Jaisson P, editors. Evolution and Culture. Cambridge, MA: MIT Press, p 105–132.
27 Rankin DJ, Bargum K, Kokko H. 2007. The tragedy of the commons in evolutionary biology. Trends Ecol Evol 22:643–651.
28 Henrich J. 2004. Cultural group selection, coevolutionary processes and large-scale coopera- 
J Econ Behav Organ 53:3–35.
29 DeDreu CKM, Page T, Putnam L. 2006. Can second-order punishment deter ver- 
verse punishment? Exp Econ 9:265–279.
30 Niskikorakis N. 2008. Punishment and counter-punishment in public goods: can we really govern ourselves? J Public Econ 92: 
91–112.
31 Rand DG, Armao JJ, Nakamur P, et al. 2010. Anti-social punishment can prevent the co-evolution of punishment and cooperation.
J Theor Biol 265:624–632.
32 Denant-Boemont L, Masclot D, Noussair CN. 2007. Punishment, counter-punishment and sanction enforcement in a social dilemma experiment. Econ Theory 33:145–167.
33 Ertan A, Page T, Putnam L. 2009. Who to punish? Individual decisions and majority rule in mitigating the free rider problem. Eur Econ Rev 53:495–511.
34 Sefton M, Shupp R, Walker JM. 2007. The effect of rewards and sanctions in provision of public goods. Econ Inq 45:671–690.
35 Gächter S, Renner E, Sefton M. 2008. The long-term benefits of punishment. Science 322: 
1510.
36 Boyd R, Gintis H, Bowles S. 2010. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science 328:617–620.
37 Gächter S. 2012. In the lab and the field: punishment is rare in equilibrium. Behav Brain Sci 35:26–28.
38 Becker W, B junkels B. 1978. A simulation study of income tax evasion. J Public Econ 34:243–252.
39 Friedland N, Maital S, Rutenberg A. 1978. A volunteer's dilemma. Evolution 65:2725–2728.
40 Eldakar O, Wilson DS, Gorman RO. 2006. The efficient interaction of indirect reciprocity and costly punishment. Nature 444:718–723.
41 Bocci P, McFarlane L. 2003. Cyberstalking: the technology of hate. Police J 76:204–221.
42 Kilding J, Gutman D, Zeh T, et al. 2002. A neutral basis for social cooperation. Neuron 35: 
395–405.
43 Deckers W, Buhnen H, Sleeking S. 1987. The impact of public transfer expenditures on tax evasion: an experimental approach. J Public Econ 34:243–252.
44 Friedland N, Maital S, Rutenberg A. 1978. The punishment that sustains cooperation is not uncommon in the “wild.” Behav Brain Sci 35:20–21.
45 Hand E, Arora J, Nakamur P, et al. 2010. Anti-social punishment can prevent the co-evolution of punishment and cooperation. J Theor Biol 265:624–632.
46 Rand DG, Nowak MA. 2013. Human cooperation. Trends Cogn Sci 17:413–425.
47 Mathew S, Boyd R. 2014. The cost of cow- ardice: punitive sentiments towards free riders in Turkana raids. Evol Hum Behav 35: 
286.
48 Henrich J, Ensminger J, McElreath R, et al. 2010. Markets, religion, community size, and the evolution of fairness and punishment. Science 327:1480–1484.
49 Gächter S. 2012. Strong reciprocity is real, but there is no evidence that uncoordinated costly punishment sustains cooperation in the wild. Behav Brain Sci 35:45–51.
50 Mathew S, Boyd R. 2011. Punishment sustains large-scale cooperation in prestate warfare. Proc Natl Acad Sci USA 108:11375–11380.
51 Bowles S, Boyd R, Mathew S, et al. 2012. The punishment that sustains cooperation is often coordinated and costly. Behav Brain Sci 35:20–21.
52 Boehm C, Barclay HB, Dentan RK, et al. 1993. Egalitarian behavior and reverse dominance. Curr Anthropol 34:227–254.
53 Dunbar RIM. 2004. Gossip in evolutionary perspective. Rev Genet Psychol 8:100–110.
54 Feinberg M, Willer R, Stellar J, et al. 2012. The virtues of gossip: reputational information sharing as prosocial behavior. J Pers Soc Psychol 102:1015–1030.
55 Sommerfeld RD, Krebsbeek HJ, Milinski M. 2008. Multiple gossip statements and their effect on reputation and trustworthiness. Proc R Soc Lond B Biol Sci 275:2529–2536.
56 Turner MM, Mazur MA, Wendel N, et al. 2003. Relational ruin or social glue? The joint effect of relationship type and gossip valence on liking, trust, and expertise. Commun Monogr 70:129–141.
57 Adams GS, Mullen E. 2012. The social and psychological costs of punishing. Behav Brain Sci 35:15–16.
58 Smith EA. 2010. Communication and collective action: language and the evolution of human cooperation. Evol Hum Behav 28:75–84.
59 Chudek M, Henrich J. 2011. Culture-genec coevolution, norm-psychology and the emergence of human prosociality. Trends Cogn Sci 15:218–226.
60 Ostrom E. 2012. Experiments combining communication with punishment options demon- strate how individuals can overcome social dilemmas. Behav Brain Sci 35:33–34.
61 Locke J. 1690. Second treatise of Government, ed McPherson CB. Indianapolis. Hackett Publishing.