Review TRENDS in Ecology and Evolution Vol.22 No.11

Punish or perish? Retaliation and collaboration among humans

Karl Sigmund

Faculty for Mathematics, University of Vienna, Nordbergstrasse 15, 1090 Vienna; and International Institute for Applied Systems Analysis, Schlossplatz 1, 2361 Laxenburg, Austria

A spate of recent investigations on reciprocation and social enforcement in humans has brought together (and sometimes divided) economists, psychologists, anthropologists, social scientists and evolutionary biologists, in addition to neurologists and students of animal behavior. Experimental work on public goods and social incentives has addressed a wealth of questions on the emotional and cognitive (proximal) factors, and also on the genetic and cultural (ultimate) evolutionary mechanisms involved in this essential aspect of human nature.

Introduction: reciprocation and social enforcement

How do humans manage to sustain collective efforts in sizable groups of unrelated individuals? This topic is in fashion but not new. In 1975, for instance, W.D. Hamilton closed his essay on 'Innate Social Aptitudes of Man' [1] with a section on 'Reciprocation and Social Enforcement'. Humans have a special gift for reciprocation. However, in interactions involving more than two individuals, reciprocation works less well than in pairwise encounters. Even defining it is a non-trivial task. If your group contains a cooperator and a cheater, whom do you reciprocate with? Social enforcement, by contrast, works better in groups with more than two members, as Hamilton points out [1], and can offer 'at least a partial cure' for the problems with reciprocation in larger groups: 'There may be reason to be glad that human life is a many-person game and not just a disjoined collection of two-person games' [1].

Whereas pioneers of sociobiology were aware of the importance of public goods and punishment [1–3], recent cross-disciplinary contact between experimental economists and evolutionary biologists has greatly stimulated the field. Here, I review work focusing on incentives which promote cooperation in groups of unrelated humans.

Fining free-riders

Let us begin with an experimental 'public good' game. Six anonymous players are given $10 each. They must decide whether to invest this in a common pool, knowing that the experimenter will triple the amount in the common pool, and distribute it equally among all six players, irrespective of whether they contributed.

This game is easy to analyze. If all players contribute, they triple their endowments. However, each player is better off by not contributing because only half of the contribution returns to the account from which it came (i.e. it is multiplied by three and then divided among the six players). If, as a consequence, no player contributes, then the initial amounts remain unchanged. This deplorable outcome of selfish motives is known variously as social dilemma, tragedy of the commons, free-rider problem or market failure. The multiplicity of the names points to the ubiquity of the issue.

In modern society, the exploitation of collective efforts (e.g. free-riding on public buses, shirking tax or dodging military service) is punished by a plethora of institutions. Obviously, the threat of punishment deters would-be defectors. This can be mimicked by another experiment on public goods, this time with punishment. In this two-stage game, the first stage runs exactly as before. In the second stage, players can impose fines upon their coplayers. These fines are collected by the experimenter and do not land in the account of the punisher. In fact, each punisher must pay a fee for the experimenter to collect the fine.

Again, the analysis is easy. A player bent on maximizing income should not punish because this is costly. Hence, nothing should happen in the second stage; thus, the first stage will be unaffected. No punishment, no contributions and no gains: the selfishly motivated inertia in both stages of the game leads to economic paralysis.

Gratifyingly, this does not happen in real experiments, which are usually slightly more sophisticated versions in which players can choose between different levels of contribution and sizes of fines. In seminal experiments by Fehr and Gächter [4,5], the average contribution of players, in the public good game without punishment, was slightly >50% of their endowment. In the public good game with punishment, it was higher – close to 60%. Punishment was usually targeted on defectors, and its mere threat had an immediate effect. However, the full size of this effect only shows when the game is repeated for several rounds (Figure 1). In the absence of punishment, contributions decrease; with punishment, they quickly increase to almost 100%. This happens if the groups stay together but, most significantly, even if the groups are newly formed between rounds, and players know that they will never meet a coplayer twice. By inflicting punishment, they can conceivably turn a defector into a cooperator. However, punishers know that the future contributions of such a 'reformed' player will exclusively benefit others. Punishment seems to be an altruistic act.

Corresponding author: Sigmund, K. (karl.sigmund@univie.ac.at).
Available online 25 October 2007.

 0169-5347/$ – see front matter doi:10.1016/j.tree.2007.06.012

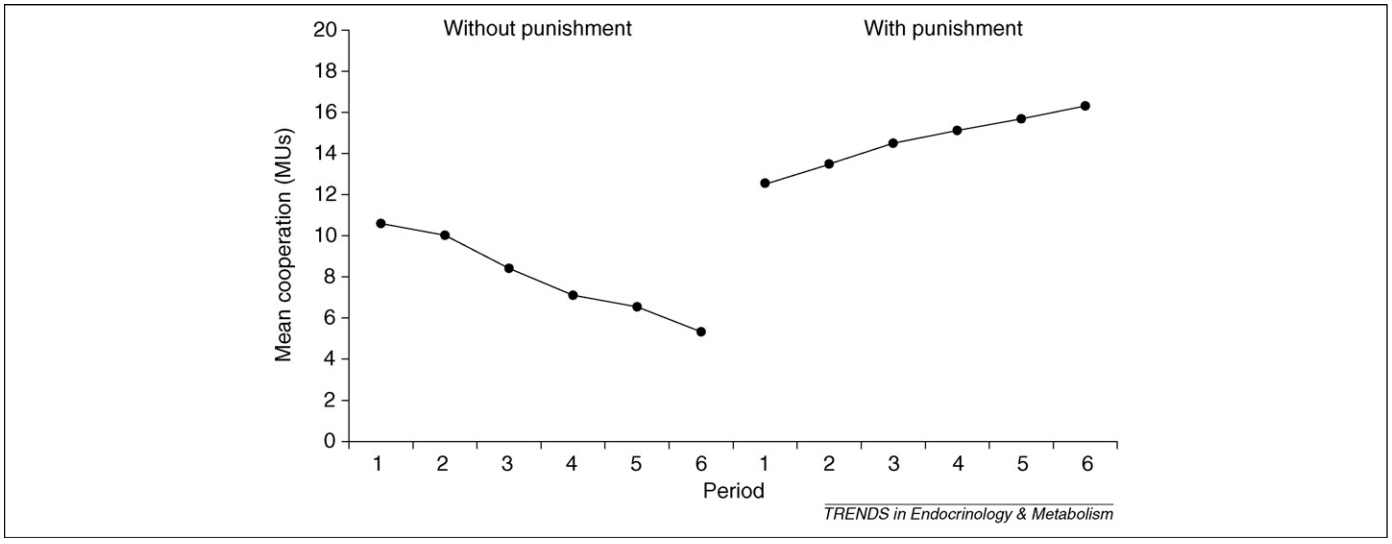

Figure 1. Learning to exploit versus teaching to cooperate. In the Fehr–Gächter experiment [4], groups of players engage in six rounds of public good without punishment, followed by six rounds with punishment. Shown are the average contributions per round. The groups are newly formed between rounds, so that players never interact with a coplayer twice. It should be emphasized that in the rounds with punishment, the average income is usually below that without punishment: punishment is costly. However, in later rounds, when most players cooperate, punishment should be rare. Ideally, it is no longer needed, except as a threat. This should yield a stable and economically efficient collaboration. Reproduced, with permission, from Ref. [4].

This is a stunning outcome. Without sanctions, the public good (i.e. the tripling of the endowment) is not realized. With sanctions, it is – although selfish reckoning prescribes that costly punishment should not be delivered. In the absence of institutions, players are willing 'to take the law into their own hands' (also known as 'peer-punishing'). This enforces cooperation in many-player interactions between unrelated individuals, which is a remarkable trait of human societies, and surely an essential factor in our evolutionary history (Figure 2).

Sanctions and social dilemmas

The investigation of the interplay between mutual assistance and social enforcement is a booming enterprise.

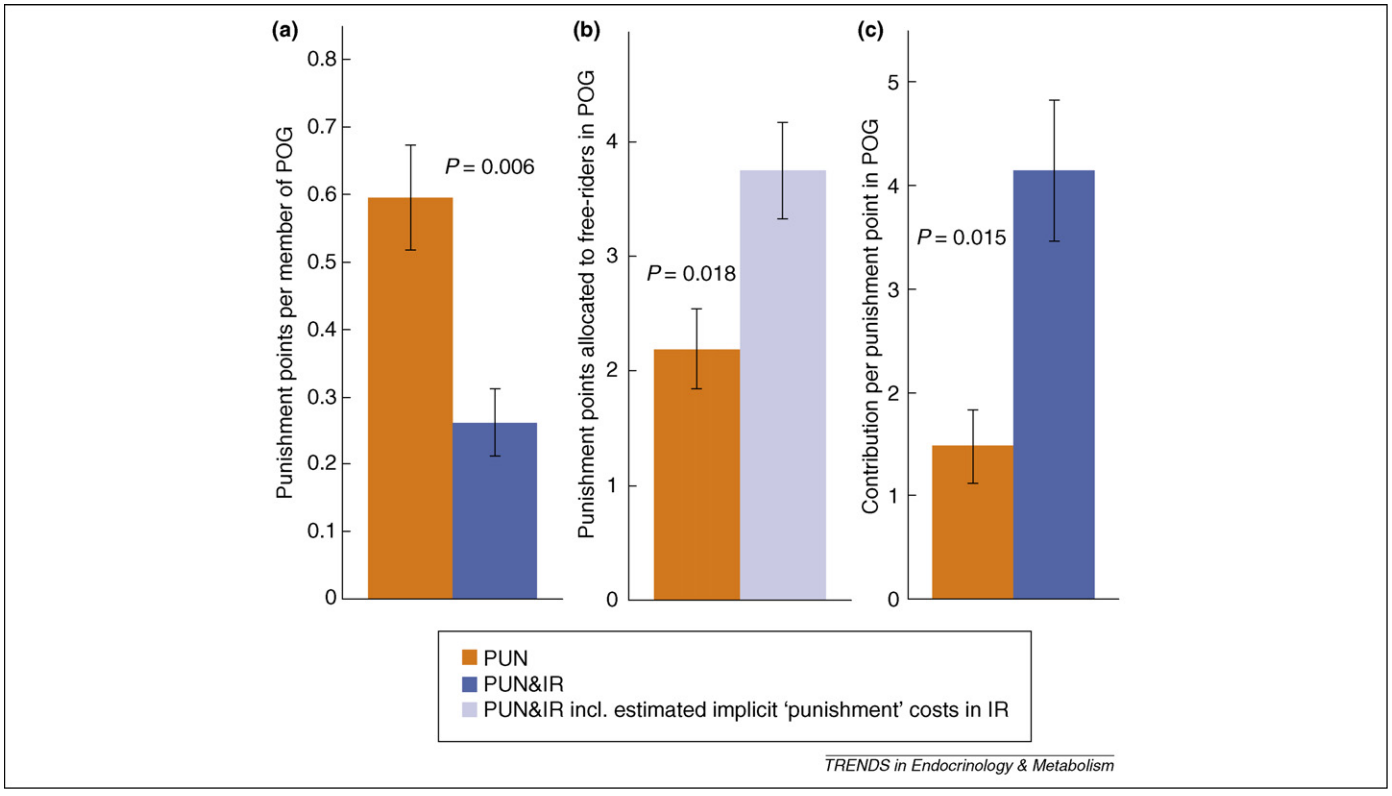

Figure 2. Peer punishment versus reputation building. In the PUN treatment, players can punish their coplayers after every round of a public good game. In the PUN&IR treatment, they engage, in addition, in several rounds of indirect reciprocity (Box 1) after each round of the public good game. Such rounds enable the rewarding of players who have contributed to the public good. By withholding a possible donation in the indirect reciprocity game, players can effectively sanction free-riders without paying a cost. This reduces the amount of direct, costly punishment **(a)** but does not eliminate it. Rather, the direct punishment is now more focused towards free-riders **(b)**, and is considerably more efficient in boosting contributions **(c)**. Reproduced, with permission, from Ref. [74].

Economists use experimental games to study the effects of positive and negative incentives (i.e. reward and punishment) on our propensity to collaborate [6,7]; anthropologists visit small-scale societies to measure the culture dependence and universality of norms that enforce cooperation [8]; psychologists study the often subconscious cues eliciting emotions that lead to helping behavior or moralistic aggression [9–11]; neurologists use magnetic resonance techniques to correlate social dilemmas with brain activities [12,13]; game theorists modify their utility functions to take account of non-monetary concerns [14,15]; biologists look for signs of policing and sanctions in bees or bacteria [16,17]; and political scientists attempt to improve the governance of institutions promoting collective actions [18,19]. Transdisciplinary dialogues are in full swing, although communication sometimes needs improving [20].

The underlying questions concerning political creatures and human nature go back at least to Aristotle. The formal framework for discussing social dilemmas that arise with public goods was provided by game theory [21]. As Olson [22] stated in 1964, in his '*Logic of Collective Action*', self-interested individuals will not behave so as to achieve their group interest, except when prodded by incentives directed selectively towards individuals in the group (i.e. punishing exploiters or rewarding contributors). In 1965, the biologist Hardin [23] addressed the same issue in a highly influential *Science* paper on the 'Tragedy of the commons', and offered as a solution to the social dilemma: 'mutual coercion, mutually agreed upon'. This is also advocated in Hobbes' *Leviathan*. But how can the agreement be enforced? The role of sanctioning institutions, in our civilization and in other societies, to uphold social norms and protect public goods has become an object of intense scrutiny [24].

Providing selective incentives for collective action is itself a public good, of course, so that the prevalence of punishing (or rewarding) institutions seems to present a chicken-and-egg problem. 'The provision of a sanctioning system as a public good' was the title of an experimental paper by Yamagishi [25], studying the effect of costly punishment on contributions towards a collective benefit. In the 1990s, a wide range of studies investigated punishing and rewarding, usually in two-person games (Box 1).

The basic finding in all of these economic experiments was that most humans are not self-centered but are other-regarding [26,27]. Their aim is not uniquely the maximization of their income. They are strongly motivated by emotions ('moral sentiments', in the words of Adam Smith).

Witness, for example, the ultimatum game (Box 1): a responder uniquely interested in income maximization should accept any positive offer, even the smallest. A similarly disposed proposer should, therefore, make the minimal offer, and keep the rest of the sum. However, the real outcome is vastly different. Most offers are close to a fair split; the rare unfair offers are mostly rejected. Dozens of experiments verified the robustness of this outcome. In particular, in all 'modern' societies, some two-thirds of the offers are between 40% and 50% of the total sum; those <20% are few, and are usually rejected [28]. A vast collective effort of anthropological studies [8] was able to document cultural variation in small-scale societies; but even among the Machiguenga, an Amazonian population of hunter-gatherers, the mean offer was 26%; this record of unfairness is still a long way from the minimal offer predicted for selfish agents.

The ultimatum game is played in a group of two players only and seems, at first glance, to be distinct from the public goods game with punishment: the players are in different roles, and each one has only one decision to make. However, the rejection of an offer by the responder is a costly punishment (less costly to the punisher if the offer is small, and hurting the proposer all the more).

The economic experiments displaying 'human nature' motivate both social psychologists and evolutionary biologists (in addition to evolutionary psychologists and sociobiologists) to study the proximal or ultimate causes (i.e. the how and the why of human cooperation). This review can only give pointers to a literature drawing, in each field, on a rich tradition.

Box 1. Game Zoo: a brief lexicon of two-person games

Many experimental two-person games are related to the issues of public good [28]. Typically, the players are anonymous, and are endowed with a certain amount of money beforehand (e.g. a show-up fee). They are asked to make their decision after having understood the rules of the game and being assigned to the roles of proposer and responder (or donor and recipient).

Gift giving: in some sense, an atom of social interaction. The donor decides whether to pay \$1 to confer a benefit of \$3 on the recipient.

Prisoner's dilemma: the mother of all cooperation games is played in many variations. In one particularly transparent setup, both players engage in a gift-giving game with each other. When players decide simultaneously, this is similar to a two-player public good game. If both cooperate by sending a gift to the other, both gain \$2. But sending a gift costs \$1, so that the best reply to whatever the coplayer decides is not to cooperate (i.e. to defect). If both players defect, however, they gain nothing.

Ultimatum: the experimenter assigns a certain sum, and the proposer can offer a share of it to the responder. If the responder (who knows the sum) accepts, the sum is split accordingly between the two players, and the game is over. If the responder declines, the experimenter withdraws the money. Again, the game is over; but this time, none of the two players gets anything.

Dictator: same as the ultimatum game, except that the responder cannot reject the offer.

Trust: in a first stage, the proposer can confer a certain benefit on the responder, as in the gift-giving game. In the second stage, the responder can decide how much of it to return to the proposer. This is similar to the sequential prisoner's dilemma game (when first one player acts as donor, and then the other).

Repeated prisoner's dilemma: the two players interact for several rounds of the prisoner's dilemma. Usually, they are not told beforehand when the interaction will be over, so as to avoid 'last round effects' (defection motivated by the fact that the coplayer cannot retaliate).

Indirect reciprocity: in a large population of players, two players are sampled at random and play the gift-giving game or the (non-repeated) prisoner's dilemma game. This is repeated again and again. The players know that they interact only once, so that retaliation is impossible.

Ultimate reasons of costly punishment

Altruistic behavior and selfish genes provide a favorite playground for theories on the evolution of cooperation, and have led to a rich toolbox (Box 2). Does this toolbox

Box 2. Tools for fitness: a semantic guide to the evolution of cooperation

Punishing and rewarding are responses to previous actions. In principle, reciprocation can only occur as the second stage of an interaction; however, the first stage of the interaction (the contribution to the public good, or the offer in the ultimatum game) is often affected by the expectation of a return. The challenge is to explain that return (i.e. the second stage).

Reciprocity operates even with third parties. This is called 'indirect reciprocation' and comes in two flavors. A bystander watching Joe harm Bill (or help Bill) is likely to harm (or help) Joe, in turn: this is vicarious reciprocity. Conversely, an individual who has been harmed, or helped, by some agent Ted, can vent his anger, or his gratitude, on some passerby, Bob: this is misdirected reciprocity. These effects have been documented in experiments [64,65]. For instance, the propensity of a bystander to punish defectors in a prisoner's dilemma game is high, except if both players defected and, thus, in a sense, performed the punishment themselves.

The emotionally driven disposition to return good with good, and bad with bad, is called 'strong reciprocity' by a group of researchers who cooperated in a vast effort to study economic games in small-scale societies. Despite substantial differences in their views, they are often perceived as an in-group around the banner of strong reciprocity [44].

Reciprocation is usually costly, whether it is altruistic (rewarding) or spiteful (punishing). Evolutionary biologists have to understand its adaptive value, and show that such behavior is (on average) not fitness-reducing after all. The classical approaches invoke interactions between kin, or mutual benefits to cooperating individuals [20]. Both approaches rely on positive assortment between individuals conferring help. The explanation is genetic in one case, and economic in the other.

Kin selection operates if a loss in direct fitness (their own reproductive success) is compensated by a gain in indirect fitness (the reproductive success of related individuals) [66,67]. Mutual benefits to cooperating individuals can accrue, for instance, if the same two players engage in a long chain of give-and-take, as with direct reciprocity, or if reputation effects enable cooperators to channel benefits towards those individuals who are benefiting others, as with indirect reciprocity. Costly signaling can be subsumed under this heading, if players who are able to signal higher value (because they can afford to contribution) are preferentially chosen as partners or mates and, thus, obtain benefits in return.

Box 3. Punishing logic: the evolutionary problems raised by costly punishment

Punishers raise two evolutionary riddles. They cannot invade; and they can be invaded.

A minority of punishers invading a population of defectors would have to punish left and right. Because each act of punishment is costly, punishers would suffer heavily, whereas the defectors would barely be affected. Hence, punishers would be at a disadvantage, and would soon be eliminated from the population.

Conversely, suppose that a population is dominated by punishers. Defectors, in this case, cannot invade: a minority of defectors would have to bear the full brunt of punishment from the majority, which more than offsets their gain from not contributing. This is a bi-stable situation: defectors cannot invade punishers, and punishers cannot invade defectors. However, suppose that a new type enters the population: one who contributes but does not punish. Such a type can easily arise through recombining traits. The newcomers do just as well as the resident punishers and, thus, can slowly spread by neutral drift. In fact, if occasionally some defectors entered the population (to be promptly eliminated by the punishers), the new type would do better than the punishers, by economizing on the cost of punishment. This new type is a second-order exploiter, free-riding on the sanctions provided by the punishers. Hence, it will spread: and this means that, eventually, there will be too few punishers to keep the defectors at bay. Thus, second-order exploiters sabotage the enforcement of contributions to the public good game in the long run and, therefore, sap the basis for the punishers, and for themselves: both contributing types will be displaced by defectors.

A remedy coming to mind is 'second-order punishment' (i.e. in punishing not only the 'first-order exploiters' who fail to contribute, but also the 'second-order exploiters' who contribute but fail to punish). However, this could give rise to 'third-order exploiters', and so on. If punishers of a sufficiently high order dominate the population, there will be few defectors and, hence, few occasions for lower-order punishers to reveal their limitations to their fiercer brethren. Thus, they would rarely be selected against, and could spread by neutral drift, sapping the system. Clearly, higher-order punishers cannot gain a foothold in a population of defectors.

offer an explanation for our propensity to punish cheaters in public good interactions? How can the trait emerge, and how can it be maintained (Box 3)?

Two evolutionary approaches to these questions are based on group selection, and invoke selective group extinction [29,30]. It seems likely that intergroup conflict was frequent in early human history, and had a large impact on shaping human instincts. If a group is threatened with extinction, solidarity soars as people can tell who have experienced bombing raids. Courage, comradeship, bonding and the readiness to risk, or sacrifice, one's own life for the group must have been shaped by such recurrent episodes from the past. It seems less compelling that this also molded the behavior observed in economic games, which is based on common concerns for fairness or reciprocation.

Several other models [31–33] exploit the fact that if populations are not well mixed but interact only locally, benefits through punishment are easier to achieve. According to one view, which defines relatedness by statistical correlation rather than common descent, both group selection and localized interaction can be translated into a kin selection framework [34]. However, as stressed by Gardner and West [35], the relatedness of social partners counts less than the effect of the punishment dealt out by an individual on the cooperation received by that individual, and this can be due to facultative adjustment, especially in small groups.

In a different approach [36], costly punishment is explained by two additional factors: (i) a tendency to copy not only the most successful strategy, but also the most frequent; and (ii) 'second-order punishment' directed at those who contribute but fail to punish. Whether 'higher-order punishment' really occurs is a moot point because experiments have failed to show evidence for it [37]. However, we surely are conformists; yet this tendency to swim with the majority, which helps to stabilize a widespread trait, works against its gaining a foothold in a population dominated by another strategy.

A similar problem besets another model [38], which assumes that, with some small probability, players defect when they learn that their coplayers do not punish. This requires information about the others, and a dose of opportunism. Again, a population of punishers cannot be invaded by exploiters, whether first or second order. But conversely, punishers are unable to invade a population of defectors. Thus, the emergence of punishers remains an open issue.

Fowler [39] suggests a possible solution, which exploits the fact that, in many public good interactions, players are not obliged to participate. Imagine that players randomly sampled from the population are offered to participate in a public good game, or to stand aside. Those who participate, and find themselves in a group of cooperators, will increase their payoff; those who participate, but land in a group of defectors, will lose. The collective effort is thus a speculation whose success depends on the coplayers. This model leads to the emergence of costly punishment, provided that a group of two or more cooperators does better than the non-participants [40,41]. If second-order exploiters manage to spread, defectors quickly take over and make the joint effort unattractive. However, such episodes are rare and short; when only a few are willing to participate in the public good game, cooperation and costly punishment reappear, and dominate most of the time. With other things being equal, a voluntary public good game with punishment is more likely to prosper than is a compulsory one.

Once punishers have invaded and taken over, all factors mentioned earlier (conformism, reputation, etc.) can join in and stabilize the propensity to punish. Individual adaptation, in this model, is based on imitation rather than on inheritance. Similarly, the spread from one group to another is easier to conceive as cultural, rather than genetic. However, once punishment is established, genetic selection will favor suitable cognitive or emotional adaptations [42].

These optimistic lines do not imply that all problems with costly punishment are solved. There are public good situations that require other models – for instance, when participation is compulsory. The substantial percentage of unconditional defectors (or cooperators) remains unexplained and so on.

Although group selection is not the only alternative, it is favored by some experimentalists [15,43] because games with anonymous subjects eliminate all possible effects of relatedness, reputation, future interactions or signaling. Hence, neither direct nor indirect reciprocity, nor kin selection nor costly signaling is at work and, thus, the reason for costly punishment has to be the only alternative remaining, namely group selection. This argument is in the venerable tradition of Sherlock Holmes: 'When you have eliminated the impossible, whatever remains, however improbable, must be the truth'. However, the experiments also eliminate group benefits [44]. This would be different if the players were told, for instance, that their group was one of several, and that the members of the group with smallest total payoff would lose all of their earnings (a setup that is likely to promote cooperation even without punishment); but players know that they are an anonymous sample from a large population, and will disperse after one round of the game. Yet such anonymity is a highly artificial condition, and many doubt that humans have evolved suitable adaptations.

As Burnham and Johnson write [44]: 'People may behave as if they are from the same evolutionarily relevant group, but in that case we may just as well assume that anonymous subjects behave as if they are related, or as if they are destined to meet again, or as if they are observed by others'. Thus, kin selection, direct or indirect reciprocity, costly signaling, in addition to group selection, might all work. Nothing can be excluded out of hand. It could even be that, as in Agatha Christie's '*Murder on the Orient Express*', all suspects collude. Theoreticians tend to look for the most parsimonious explanation, but this principle need not be appropriate to the historical contingencies of human evolution.

Proximate causes of costly punishment

It makes no sense to assume that ultimatum games or public good games, in their clinical sterility, have shaped our evolution, although human behavior in these games is based on evolved traits. The stark artificiality of economic experiments helps (as in physics or physiology) to reveal the mechanisms underlying these traits.

It seems from cross-cultural studies that the readiness to inflict costly punishment on cheaters is a human universal [8]. It varies across societies but is strongly correlated with altruistic behavior, such as the readiness to help others in dictator games. Brain imaging techniques show that when players inflict costly punishment, special satisfaction-related zones in the dorsal striatum are activated, indicating physiological adaptations [13]. Both punishing and rewarding seem to be facets of the deep-seated human propensity to reciprocate good and bad, a propensity guided by reasoning and emotions, and based on heuristics and cues (see Box 4).

Two recurrent findings of experimental economics are, on the one hand, the diversity within populations, and, on the other hand, the flexibility of individuals. All populations seem to be polymorphic, with a substantial percentage displaying little reciprocation. If groups of 'high trusters' or 'low trusters' are assorted according to simple test questions, they achieve different levels of cooperation in public good games [25]. However, many humans can adapt quickly and fine-tune their actions to their social environment. Players do not merely respond to the threat of punishment or the promise of a reward but they update constantly, taking account of their experience [5]. If players are told that they will be rematched with the same coplayers, or that their decisions will be made known, they often change their behavior, obviously motivated by concerns for longer-lasting interactions or for reputation [45–48]. Similarly, if they can opt out of the public good game, or back into it, they base their decisions on the current state of the population and adapt rapidly [49]. Voluntary participation elicits a greater readiness to cooperate [50] (Box 4).

This alertness can misfire: it has been shown that by merely seeing the image of an eye, players can be motivated to increase donations substantially [51–53]. Such obvious maladaptations strongly support the hypothesis that our evolutionary legacy shapes our economic behavior. Similarly, cooperation can be increased by cues of reciprocity or kinship [54]: a face with a family resemblance elicits more help. It is also well known that seemingly unimportant factors (for instance, a preference for Klee rather than Kandinsky) can establish a group identity among complete strangers and boost solidarity [55]. Even in the absence of cues, players could be influenced by the relevant concerns, at least in the sense of hedging their bets. (The tendency to

Box 4. The carrot: the role of rewards as incentives for cooperation

Investigations comparing negative with positive incentives (i.e. the carrot with the stick) show that rewards are considerably less efficient than punishment, at least for the games considered here, in which the public good is a linear function of the number of contributors. Positive incentives become costly, and negative incentives become cheap, if success is fully achieved – that is, all cooperate [10,68–70].

Andreoni *et al.* [71] studied four treatments of a proposer–responder game in which the proposer had to choose how much to share of a given sum. Depending on the treatment, the responder subsequently had the possibility: (i) to reward or to punish the proposer; (ii) only to reward; (iii) only to punish; or (iv) neither to reward nor to punish. Treatment (iv) becomes the same as the dictator game (Box 1); treatment (iii) with the punishing option differs from the ultimatum game (Box 1) because the responder has more freedom in choosing the proper sanction.

Rewards alone prove ineffective. Punishment, by contrast, often induces offers close to 50%. Adding the possibility of rewarding yields a remarkable outcome: half of the offers are >50%, and more than a quarter of the proposers offer 100%. The corresponding reward is half of that. Punishment is hardly affected by the availability of rewards but rewarding is considerably more pronounced if there is no possibility of punishment.

A particularly interesting system of incentives was considered by Milinski *et al.* [72] and Panchanathan and Boyd [73]. Between rounds of the public good game without punishment, the population engages in pairwise interactions of indirect reciprocity (Box 1). Because players tend preferentially to help those who have contributed to the public good, this effectively provides rewards which are not costly. This is because those who reward earn a good reputation and thus benefit in later rounds of the indirect reciprocity game. If, in addition to indirect reciprocity, the players have an opportunity directly to peer-punish those who do not contribute, they use this opportunity less often but in a more focused way [74] (Figure 2). Costly punishment and rewarding through indirect reciprocity combine efficiently to boost cooperation in group interactions.

invest roughly half in the first round of a public good game could be such an insurance policy.)

Our understanding of when and why subliminal factors can affect decision making is far from complete. Players can strongly react to an appropriate cue, even when knowing that reality does not back it up. (An example often mentioned in discussions is the sexual arousal produced by centrefolds.) In particular, the fear of punishment can be easily evoked. Under normal circumstances, the donations in the dictator game (Box 1) are smaller than in the ultimatum game because proposers understand that low offers cannot be rejected. However, if proposers know that they will be informed of what the responder thinks of their offer, they offer as much as with the ultimatum game [56]. Such purely symbolic punishment is no longer costly. It has been argued that a strong motive for cooperation and moral behavior is the fear of punishment by supernatural spirits [57]. Superstitious maladaptations are widespread, possibly because they strongly promote conformism and obedience.

Fear, shame, guilt and their converse, the elation and inner glow after a generous action, work to keep humans from cheating. Being cheated arouses anger, indignation and moral outrage, and often causes individuals to inflict costly punishment on defectors.

However, reducing punishment to retaliatory motives and anger at norm-breakers might be premature. Recently, the public good experiment with punishment was repeated, with the difference that the first stage (public good) was replaced by a lottery [58]. Players received randomly assigned sums (distributed as in the Fehr–Gächter experiment [4]) and then could inflict costly 'punishment' just as before, except that they were fully aware that their coplayers had done nothing wrong. Many chose to reduce the top earners' income, producing an effect statistically undistinguishable from the reduction of the income of below-average contributors in the public good game with punishment. Inequality arouses negative emotions.

If the public good game is repeated, but this time with a fee:fine ratio of 1:1 (the punisher has to pay as much as the punished), then the difference between the two players' payoffs is not altered by punishment (although payoff variance can be reduced). Nevertheless, contributors punish defectors vigorously, and the threat of punishment boosts contributions [59]. Significantly, whereas in the 1:3 treatment defectors sometimes impose sanctions (on defectors and cooperators alike), this rarely happens in the 1:1 treatment. Defectors seem to be little affected by fairness norms. By contrast, cooperators (who usually punish only defectors) more than double their efforts on imposing fines, obviously willing to incur higher costs to inflict the 'just' retribution on wrong-doers.

The limitations of peer-punishment

Although punishment works to boost cooperation, it can also be counterproductive. It often lowers the average income in public good games, despite raising the average level of contributions. In games of trust, or games involving rewards, adding the threat of punishment can decrease the menaced player's willingness to cooperate [60]. In a particularly elegant set of experiments, it has been shown that, if players of a public good game are offered before each round the choice between the versions with or without punishment, many tend first to shun negative incentives. They need a few rounds to learn to switch to the version with sanctions [61]. Together with the theoretical model of a public good game with punishment, based on voluntary participation [41], this provides a sound application of Hardin's principle 'Mutual coercion, mutually agreed upon' [23].

Punishment is not the only way to enforce cooperation; harassing those having access to a resource [62], chasing shirkers [63] or sabotaging the attempts of cheaters [16] are different examples, and can also be found in other animals, such as mammals, fish or insects. However, humans, with their cognitive capacities for individual recognition, temporal discounting, memory, empathy and language, are uniquely gifted to develop the proximal mechanisms needed for reciprocation, and in particular for punishment.

Yet it needs to be stressed that peer punishment seems to be relatively rare in real life (in contrast to experiments under anonymity). It can be costly indeed. In small-scale societies, or village life, reputation might have a more pervasive role. It is easier to gossip behind the back of a bully than to confront him. Undermining a good reputation is an inexpensive but ominous form of sanctioning, which

might eventually lead to ostracism – that is, exclusion from the market for trustworthy partners. In large societies, peer punishment is also rare, and is repressed by the institutions upholding law and order. Both the pervasive market economy for reliable partners and the step from peer punishment to the establishment of sanctioning institutions deserve closer future investigation.

Acknowledgement

I would like to thank Hannelore Brandt, Robert Boyd, Daniel Fessler, Simon Gächter, Manfred Milinski, Mayuko Nakamaru, Martin Nowak and Bettina Rockenbach for helpful discussions. This work was funded by EUROCORES TECT I–104–G15.

References

1 Hamilton, W.D. (1996) *Narrow Roads of Geneland, Collected Papers I,* Freeman
2 Trivers, R.L. (2002) *Natural Selection and Social Theory: Selected Papers of Robert Trivers,* Oxford University Press
3 Clutton-Brock, T.H. and Parker, G.A. (1995) Punishment in animal societies. *Nature* 373, 209–216
4 Fehr, E. and Gächter, S. (2002) Altruistic punishment in humans. *Nature* 415, 137–140
5 Fehr, E. and Gächter, S. (2000) Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* 90, 980–994
6 Charness, G. and Haruvy, E. (2002) Altruism, fairness and reciprocity in a gift-exchange experiment: an encompassing approach. *Games Econ. Behav.* 40, 203–231
7 Masclet, D. *et al.* (2003) Monetary and non-monetary punishment in the voluntary contributions mechanism. *Am. Econ. Rev.* 93, 366–380
8 Henrich, J. (2006) Costly punishment across human societies. *Science* 312, 176–177
9 Price, M.E. *et al.* (2002) Punitive sentiment as an anti-free rider psychological device. *Evol. Hum. Behav.* 23, 203–231
10 Baumeister, R.F. *et al.* (2001) Bad is stronger than good. *Rev. Gen. Psychol.* 5, 323–370
11 Kurzban, R. and Houser, D. (2005) An experimental investigation of cooperative types in human groups: a complement to evolutionary theory and simulations. *Proc. Natl. Acad. Sci. U. S. A.* 102, 1803–1807
12 Sanfey, A.G. *et al.* (2003) The neural basis of economic decision making in the ultimatum game. *Science* 300, 1755–1758
13 de Quervain D.J-F. *et al.* (2004) The neural basis of altruistic punishment. *Science* 305, 1254–1258
14 Fehr, E. and Schmidt, K. (2000) A theory of fairness, competition and cooperation. *J. of Economic Perspectives* 14, 159–181
15 Fehr, E. and Fischbacher, U. (2005) Human altruism – proximate patterns and evolutionary origins. *Anal. Kritik* 27, 6–47
16 Wenseleers, T. and Ratnieks, F.L.W. (2006) Comparative analysis of worker reproduction and policing in eusocial hymenoptera supports relatedness theory. *Am. Nat.* 168, E163–E179
17 Kiers, E.T. *et al.* (2003) Host sanctions and the legume-rhizobium mutualism. *Nature* 425, 78–81
18 Gneezy, U. and Rustichini, A. (2000) A fine is a price. *J. Legal Stud.* 29, 1–17
19 Ostrom, E. and Walker, J. (2003) *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research,* Russell Sage Funds
20 West, S.A. *et al.* (2007) Social semantics: altruism, cooperation, mutualism and strong reciprocity. *J. Evol. Biol.* 20, 415–432
21 Samuelson, P.A. (1954) The pure theory of public expenditure. *Rev. Econ. Stat.* 36, 387–389
22 Olson, M. (1965) *The Logic of Collective Action,* Harvard University Press
23 Hardin, G. (1968) The tragedy of the commons. *Science* 162, 1243–1248
24 Ostrom, E. (1990) *Governing the Commons,* Cambridge University Press
25 Yamagishi, T. (1986) The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* 51, 110–116
26 Hammerstein, P. (ed.) (2003) *Genetic and Cultural Evolution of Cooperation*, MIT Press
27 Gintis, H. *et al.*, eds (2005) *Moral Sentiments and Material Interests: the Foundations of Cooperation in Economic Life*, MIT Press
28 Camerer, C. (2003) *Behavioral Game Theory: Experiments in Strategic Interactions,* Princeton University Press
29 Boyd, R. *et al.* (2003) The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3531–3535
30 Bowles, S. and Gintis, H. (2002) Homo reciprocans. *Nature* 415, 125–128
31 Brandt, H. *et al.* (2003) Punishment and reputation in spatial public goods games. *Proc. R. Soc. Lond. B. Biol. Sci.* 270, 1099–1104
32 Nakamaru, M. and Iwasa, Y. (2005) The evolution of altruism by costly punishment in lattice-structured populations: score-dependent viability vs score-dependent fertility. *Evol. Ecol. Res.* 7, 853–870
33 Nakamaru, M. and Iwasa, Y. (2006) The evolution of altruism and punishment: role of the selfish punisher. *J. Theor. Biol.* 240, 475–488
34 Lehmann, L. et al. Strong reciprocity or strong ferocity? A population genetic view of the evolution of altruistic punishment. *Am. Nat.* (in press)
35 Gardner, A. and West, S.A. (2004) Cooperation and punishment, especially in humans. *Am. Nat.* 164, 753–764
36 Henrich, J. and Boyd, R. (2001) Why people punish defectors. *J. Theor. Biol.* 208, 79–89
37 Kiyonari, T. *et al.* (2004) Second-order punishment in one-shot social dilemma. *Int. J. Psychol.* 39, 329
38 Sigmund, K. *et al.* (2001) Reward and punishment. *Proc. Natl. Acad. Sci. U. S. A.* 98, 10757–10762
39 Fowler, J.H. (2005) Altruistic punishment and the origin of cooperation. *Proc. Natl. Acad. Sci. U. S. A.* 102, 7047–7049
40 Brandt, H. *et al.* (2006) Punishing and abstaining for public goods. *Proc. Natl. Acad. Sci. U. S. A.* 103, 495–497
41 Hauert, C. *et al.* (1905-1907) Via freedom to coercion: the emergence of costly punishment. *Science* 316, 1905–1907
42 Richerson, P. and Boyd, R. (2005) *Not By Genes Alone,* University of Chicago Press
43 Fehr, E. and Henrich, J. (2003) Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. In *The Genetic and Cultural Evolution of Cooperation* (Hammerstein, P., ed.), pp. 55–82, MIT Press
44 Burnham, T. and Johnson, D.D.P. (2005) The biological and evolutionary logic of human cooperation. *Anal. Kritik* 27, 113–135
45 Sell, J. and Wilson, R.K. (1999) The maintenance of cooperation: expectations of future interactions and the trigger of group punishment. *Soc. Forces* 77, 1551–1570
46 Small, D.A. and Loewenstein, G. (2005) The devil you know: the effects of identifiability on punishment. *J. Behav. Dec. Mak.* 18, 311–318
47 Rege, M. and Telle, K. (2004) The impact of social approval and framing on cooperation in public good situations. *J. Public Econ.* 88, 1625–1644
48 O'Gorman, R. *et al.* (2005) Altruistic punishing and helping differ in sensitivity to relatedness, friendship, and future interactions. *Evol. Hum. Behav.* 26, 375–387
49 Semmann, D. *et al.* (2003) Volunteering leads to rock-paper-scissors dynamics in a public goods game. *Nature* 425, 390–393
50 Orbell, J.H. and Dawes, R.M. (1993) Social welfare, cooperator's advantage, and the option of not playing the game. *Am. Soc. Rev.* 58, 787–800
51 Haley, K. and Fessler, D. (2005) Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evol. Hum. Behav.* 26, 245–256
52 Burnham, T. and Hare, B. Engineering cooperation: does involuntary neural activation increase public goods contributions? *Hum. Nat.* (in press)
53 Bateson, M. *et al.* (2006) Cues of being watched enhance cooperation in a real-world setting. *Biol. Lett.* 2, 412–414
54 DeBruine, L. (2005) Facial resemblance enhances trust. *Proc. R. Soc. Lond. B. Biol. Sci.* 269, 1307–1312
55 Yamagishi, T. *et al.* (1999) Bounded generalised reciprocity: ingroup boasting and ingroup favouritism. *Adv. Group Proc.* 16, 161–197
56 Xiao, E. and Houser, D. (2005) Emotion expression in human punishment behaviour. *Proc. Natl. Acad. Sci. U. S. A.* 102, 7398–7401
57 Johnson, D. and Bering, J. (2006) Hand of God, mind of man: punishment and cognition in the evolution of cooperation. *Evol. Psychol.* 4, 219–233
58 Dawes, C.T. *et al.* (2007) Egalitarian motives in humans. *Nature* 446, 794–796

59 Falk, A. *et al.* (2005) Driving forces behind informal sanctions. *Econometrica* 73, 2017–2030
60 Fehr, E. and Rockenbach, B. (2003) Detrimental effects of sanctions on human altruism. *Nature* 422, 137–140
61 Gürerk, O. *et al.* (2006) The competitive advantage of sanctioning institutions. *Science* 312, 108–111
62 Stevens, J.R. *et al.* (2005) Evolving the psychological mechanisms for cooperation. *Ann. Rev. Ecol. Evol. Syst.* 36, 499–518
63 Bshary, R. and Grutter, A.S. (2005) Punishment and partner switching causes cooperative behaviour in a cleaning mutualism. *Biol. Lett.* 1, 396–399
64 Wedekind, C. and Milinski, M. (2000) Cooperation through image scoring in humans. *Science* 288, 850–852
65 Fehr, E. and Fischbacher, U. (2004) Third-party punishment and social norms. *Evol. Hum. Behav.* 25, 63–87
66 Frank, S.A. (2003) Repression of competition and the evolution of cooperation. *Evolution Int. J. Org. Evolution* 57, 693–705
67 Lehmann, L. and Keller, L. (2006) The evolution of cooperation and altruism – a general framework and a classification of models. *J. Evol. Biol.* 19, 1365–1376
68 Dickinson, D.L. (2001) The carrot vs. the stick in work team motivation. *Exp. Econ.* 4, 107–124
69 McCabe, K.A. *et al.* (2003) Positive reciprocity and intentions in trust games. *J. Econ. Behav. Org.* 52, 267–275
70 Sefton, M. et al. The effects of rewards and sanctions in provision of public goods. *Econ. Inq.* (in press)
71 Andreoni, J. *et al.* (2003) The carrot or the stick: rewards, punishments, and cooperation. *Am. Econ. Rev.* 93, 893–902
72 Milinski, M. *et al.* (2002) Reputation helps solve the 'Tragedy of the Commons'. *Nature* 415, 424–426
73 Panchanathan, K. and Boyd, R. (2006) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* 432, 499–502
74 Rockenbach, B. and Milinski, M. (2006) The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444, 718–723