Punishment is not a group adaptation

Humans punish to restore fairness rather than to support group cooperation

Nicolas Baumard

Institute of Cognitive and Evolutionary Anthropology
University of Oxford

Institut Jean-Nicod
CNRS-EHESS-ENS

nbaumard@gmail.com

Abstract

Punitive behaviours are often assumed to be the result of an instinct for punishment. This instinct would have evolved to punish wrongdoers and it would be the evidence that cooperation has evolved by group selection. Here, I propose an alternative theory according to which punishment is a not an adaptation and that there was no specific selective pressure to inflict costs on wrongdoers in the ancestral environment. In this theory, cooperation evolved through partner choice for mutual advantage. In the ancestral environment, individuals were in competition to be recruited in cooperative ventures and it was vital to share the benefits of cooperation in a mutually advantageous manner. If individuals took a bigger share of the benefits, their partners would leave them for more interesting partners. If they took a smaller share, they would be exploited by their partners who would receive more than what they had contributed to produce. This competition led to the selection of a sense of fairness, a cognitive adaptation aiming to share equally the benefits of cooperation in order to attract partners. In this theory, punishment is not necessary for the evolution of cooperation. Punitive behaviours are only a way to restore fairness by compensating the victim or penalizing the culprit. Drawing on behavioural economics, legal anthropology, and cognitive psychology, I show that empirical data fit better with this framework than with the theory of group selection. When people punish, they do so to restore fairness rather than to help the group.

1. Introduction

For a long time, evolutionary theories of human cooperation were dominated by 'mutualistic' theories such as reciprocal altruism (Trivers, 1971) or indirect reciprocity (Alexander, 1987) in which cooperation arises because it is mutually beneficial to individuals (West, Griffin, & Gardner, 2007). The observation of punitive behaviours in behavioural games (Fehr & Gächter, 2002) has demonstrated the limits of standard mutualistic theories since in these experiments, participants are ready to punish cheaters at a cost to themselves and for no direct benefits (Gintis, Bowles, Boyd, & Fehr, 2003). By contrast, the existence of punitive behaviours seems to favour 'altruistic' theories in which individuals sacrifice for the group (Gintis, et al., 2003; Sober & Wilson, 1998). Indeed, it is exactly the kind of altruistic behaviour predicted by group selection. Furthermore, it suggests that cooperation has evolved through the sacrifice of altruistic individuals who were ready to incur some costs to prevent cheating.

In response to these altruistic approaches to punishment, supporters of mutualistic theories have proposed to explain the evolution of punishment in terms of the individual's interest in curbing the benefits of cheating (Price, Cosmides, & Tooby, 2002), in collectively controlling powerful individuals (Boehm, 2000), or in managing moral reputation (Barclay, 2006). Here, I propose a radical alternative to these responses, namely that punishment is a not an adaptation and that there was no specific selective pressure to inflict costs on wrongdoers in the ancestral environment.

This alternative is based on a contractualist approach to cooperation (Baumard, 2008, 2010b; see also Gauthier, 1986; Rawls, 1971). In this theory, cooperation evolved through partner choice for mutual advantage and consists in respecting others' interests exactly as if individuals had really agreed on a contract. In the ancestral environment, individuals were in competition to be recruited in the most fruitful ventures, and it was vital to share the benefits of cooperation in a mutually advantageous manner. If individuals took a bigger share of the benefits, their partners would leave them for more interesting partners. If they took a smaller share, they would be exploited by their partners who would receive more than what they had contributed to produce. This context is thus likely to have led to selection for a 'sense of fairness', a cognitive adaptation promoting and enabling respect for others' interests. This theory predicts the kind of morality proposed by contractualist philosophers. Individuals should behave with each other as if they had agreed on a mutually advantageous contract. In this contractualist approach, cooperation (fairness) evolves by partner choice, and does not

require any use of punishment. Thus, there is no instinct for punishment. Punitive behaviours are only a way to restore fairness by compensating the victim or penalizing the wrongdoers.

In the following sections, I detail further this theory and draw on behavioural economics, legal anthropology, and cognitive psychology to show that if offers a better theory of punitive behaviours. More precisely, I show that cooperation does not need punishment to evolve, and that when people inflict a cost on others, they do not aim to sustain group cooperation. But first, we first need to define more precisely what'punishment' is and what it is not.

2. A contractualist approach of punishment

2.1 From the market of cooperation to the sense of fairness

In the introduction, I proposed to see punishment as a logical consequence of a sense of fairness. People punish in order to compensate the victim or penalize the culprit. How has this sense of fairness evolved? I will suggest here that it is an adaptation to survive in a cooperative environment. Humans beings belong to a highly cooperative species and get most of their resources from collective actions, solidarity, exchanges, etc. (Gurven, 2004b; Hill & Kaplan, 1999). In the ancestral environment, individuals were in competition to be recruited for the most fruitful ventures and it was vital to share the benefits of cooperation in a mutually advantageous manner. If individuals took a bigger share of the benefits, their partners would leave them for more interesting partners. If they took a smaller share, they would be exploited by their partners who would receive more than what they had contributed to produce. This competition to attract cooperative partners is thus likely to have led to selection for a 'sense of fairness', a cognitive device that gives humans the intuition that others have 'rights,' 'claims,' or 'entitlements' to certain resources and that binds them to respect others' interests (goods, services, investments, etc.) in cooperative interactions (Baumard, 2008, 2010b; Baumard, Boyer, & Sperber, 2010). Humans thus behave as if they had agreed on a mutually advantageous contract with others.

A good way to understand this evolutionary theory of fairness is to start from primate cooperation. In the last two decades, primates have been shown to exchange a range of commodities such as meat, grooming, sex or political support (e.g. de Waal, 1997; Koyama, Caws, & Aureli, 2006). These exchanges can be described in terms of market (Noe & Hammerstein, 1994; Noë, van Schaik, & Van Hooff, 1991). The theory predicts that exchange rates of goods and services should fluctuate according to the law of supply and demand. In

line with market theory, many studies show that individuals pay more when commodities become scarce: Subordinates groom dominants for longer before being tolerated at food sites in periods of shortage (Barrett, Gaynor, & Henzi, 2002; Chancellor & Isbell, 2009; Port, Clough, & Kappeler, 2009); females groom mothers for longer before obtaining permission to handle their infants when there are fewer newborns (Gumert, 2007a; Henzi & Barrett, 2002) and males groom fertile females for longer before obtaining their compliance when fewer such females are present (Gumert, 2007b).

In this market of cooperation, individual have an interest in respecting the value of the commodities exchanged. If an individual offers less than the market value, then her partner will seek more profitable partners. If she offers more than the market value, she will be exploited by her partner. Individuals thus have an interest in being *fair*. They have an interest in respecting others' interests in the exchange (the value of their commodities in the market). At the psychological level, however, it is possible to respect others' interests by relying only on selfish motives because, for instance, one needs the commodity or because one wants to appear as an interesting partner. This is what non human primates apparently do: they respect others' interests when they need them, and depart from this strategy when a more advantageous alternative exists. For instance, although chimpanzees can negotiate and make concessions over the distribution of the benefits of cooperation, dominants ignore others' interests whenever possible (Melis, Hare, & Tomasello, 2009).

Non-human primates thus seem to cooperate in a cynical manner: They respect others' interests only to the extent that this strategy can improve their situation (see also Jensen, Hare, Call, & Tomasello, 2006; Silk, et al., 2005). However, this strategy may have been detrimental in humans. Indeed, the development of cooperative activities (Tomasello, Carpenter, Call, Behne, & Moll, 2005), the growing importance of cooperation in life history (Kaplan, Hill, Lancaster, & Hurtado, 2000), and the increasing size of social networks (Dunbar, 1993) may all have increased the level of competition between cooperative partners, and therefore the importance of a good reputation. This context may have led to specific cognitive adaptation. More specifically, it may have become advantageous to be *intrinsically* disposed to respect others' interests in cooperative interactions, that is to have a 'sense of fairness', and not to rely only on extrinsic and selfish motivations such as the need to get the commodity and the desire to get a good reputation.

Trivers (1971) made the same argument with respect to reciprocal altruism:

> "Selection may favor distrusting those who perform altruistic acts without the emotional basis of generosity or guilt because the altruistic tendencies of such individuals may be less reliable in the future. One can imagine, for example, compensating for a misdeed without any emotional basis but with a calculating, self-serving motive. Such an individual should be distrusted because the calculating spirit that leads this subtle cheater now to compensate may in the future lead him to cheat when circumstances seem more advantageous (because of unlikelihood of detection, for example, or because the cheated individual is unlikely to survive)." (Trivers, 1971, p. 51)

More generally, a calculating spirit may actually be a hindrance to achieving a good reputation. Many theorists have emphasized that when one wants to convince others of one's willingness to respect their possessions, directly trying to achieve a good reputation can be a bad idea, because one takes the risk of appearing to be acting with self-serving motives (Baumard & Sperber, 2007; Frank, 1988; Gauthier, 1986; Trivers, 1971). One may, for instance, talk too much about one's good deed, or appear to be morally inconsistent by being moral in one context (where friends are around) and less moral in others (where only acquaintances are around). In line with these observations, experimental research has convincingly shown that humans respond to cooperative acts according to their perception of the motives of the individual: they tend to respond more cooperatively when they perceive the other as cooperating genuinely—that is, voluntarily performing a moral act as an end in itself, without seeking any personal gain (Brehm & Cole, 1966; Krebs, 1970; Schopler & Thompson, 1968).

To sum up, it is better to be intrinsically inclined to respect the value of others possessions or contribution than to respect it for extrinsic reasons. The human disposition to cooperate may therefore be described as a *sense of fairness*, a specialised mechanism selected in an increasingly competitive market, and designed to bind individuals to respect their partners' interests. This disposition gives humans the intuition that others have some "rights," "claims" or "entitlements" on some resources. The more someone violates these rights, the more immoral her action is. Or, in market terms, the more valuable a resource is on the market, the more unfair it is to steal or destroy it.

This approach to fairness differs from mutualistic theories which defend the view that individuals respect individual interests because they care about their moral reputation (Haley & Fessler, 2005; Kurzban, 2001). The difference resides at the psychological level. For the

theory I have presented here, people are fair because, psychologically, they have a motivation to be fair. By contrast, in theories of reputation, people are fair because they care about their reputation. They do not genuinely want to be fair. They want to achieve the reputation of being a fair person.

The contractualist theory is thus a two-step theory:

Step 1: Humans have developed more and more ways to cooperate with each other, creating a market for cooperation.

Step 2: Competition among cooperative partners has led to the selection of a psychological disposition to be motivated by fairness concerns.

This account of the human cooperative disposition suggests a very straightforward evolutionary history. In primate markets, individuals already had an incentive to respect others' interests (Step 1). What humans have evolved is simply an intrinsic motivation to respect these interests (Step 2). If this view is correct, then there is therefore a genuine continuity between non-human animals and humans in this domain. Despite appearances (Hammerstein, 2003), animal cooperation and human cooperation may have followed the same mutualistic evolutionary pathway, differing only at the proximal level (scale of cooperation, psychological mechanisms).

The theory I propose here is thus a combination of two recent theories: the theory of biological markets (Noe & Hammerstein, 1994; Noë, et al., 1991) and the theory of partner choice (Bull & Rice, 1991; Peck, 1993; G. Roberts, 1998). I call it 'contractualist' since it corresponds to the classic contractualist theory of morality proposed by philosophers (Gauthier, 1986; Rawls, 1971). Indeed, in the contractualist theory, cooperation or morality is about being fair. It thus differs from classic mutualistic theories (Alexander, 1987; Trivers, 1971) in which cooperation is about helping or giving to others. In these theories, cooperative dispositions are often seen as psychological devices selected to motivate individuals to give resources to others. In the theory sketched above, however, we have seen that self-serving motives are enough to motivate individuals to give resources. The function of our cooperative disposition is thus not to *motivate exchange* (which is already achieved in Step 1 by self-serving motivations), but to *regulate exchanges*. In other words, it is not about *giving*, it is *refraining from stealing*. In other words, it is about fairness.

2.2 Fairness in human interactions

Following the primate literature, I have only considered fairness in the context of exchanges. However, the same logic can be applied to more complicated cooperative interactions. Indeed, humans cooperate in collective actions such as collective hunting, collective breeding, collective cooking, etc. A collective action can be seen as a transient cooperative venture in which partners invest some of their resources (goods and services) to produce new and more valuable resources (e.g. a food, shelter, etc.). In other words, they offer their contribution in exchange for a share of the benefits. In this situation, potential partners need to evaluate the value of each contribution on the market and to proportionate the share of the benefits to this value. If they give less than the value of the contribution, their partners will leave them for more interesting partners. If they give more, they will be exploited by their partners who will receive more than what they have contributed to produce. The contractualist theory thus predicts that partners will choose to distribute the benefits of collective actions according to the value of each individual's contribution on the market.

Experimental data accord with this prediction, showing a massive preference for meritocratic distributions: the more valuable your input, the more you get (Konow, 2003; Marshall, Swift, Routh, & Burgoyne, 1999). Similarly, field observations of hunter-gatherers have shown that hunters share the benefits of the hunt according to each participant's contribution (Gurven, 2004a). Bailey (1991), for instance, reports that in group hunts among the Efe Pygmies, initial game distributions are biased toward members who participated in the hunt, and that portions are allocated according to the specific hunting task. The hunter who shoots the first arrow gets an average of 36% , the owner of the dog who chased the prey gets 21%, and the hunter who shoots the second arrow gets only 9% by weight (see also Alvard & Nolin, 2002 for the distribution of benefits among whale hunters).

Market theory goes further in predicting that a fair distribution should be based on contributions only if these contributions have been acquired in a fair way. For instance, people clearly distinguish between fair and unfair contributions. For instance, many people think that CEO or football stars' salaries are too high compared to their contribution. Their salaries are viewed as the product of the arbitrary forces of financial markets and not as genuine and fair contributions to the collective benefit. Consequently, many people think that CEO or football players do not deserve the money they earn (for an experimental approach, see Konow, 2003). In the same way, people distinguish between fair and unfair prices. For example, Kahneman, Knetsch and Thaler (1986) report that 82% of the participants

interviewed in their study consider unfair for a hardware store to increase the price of snow shovels just after a blizzard (see also Frey & Pommerehne, 1993). In this case, the price does not reflect the contribution of the seller (he did not do any work) but rather the fact that the seller tries to take an unfair advantage of the blizzard.

The kind of collective actions we just discussed—in which the distribution of the benefits is based on individual contribution—is not the only way to cooperate. In many occasions, communal sharing or mutual aid may be more advantageous. In mutual aid, distributions are based on need as in the old Marxist maxim: "From each according to his ability, to each according to his needs!". Mutual aid may be favoured for several reasons. In highly risky activities, an equal distribution provides insurance against risks. Among hunter-gatherers, non-meat items and cultigens whose production is highly correlated with effort are often distributed according to the merit while meat items whose production is highly unpredictable are distributed much more equally (Alvard, 2004; Gurven, 2004b; Wiessner, 1996). Similarly, it is often possible in hunter-gatherers societies to distinguish the primary distribution based on merit in which hunters gets retributed for their contribution to the hunt and the secondary distribution based on need in which the same hunters share their meat with their neighbours in order to get an insurance against adversity (Alvard, 2004; Gurven, 2004b). Another reason may be that individuals (spouses, friends, kin) cooperate on a long term basis and consider that, in the long run, contributions are roughly equal (Clark & Jordan, 2002; Clark & Mills, 1979; Fiske, 1992).

The need to be recruited in cooperative networks will lead individuals to help each other in a fair way: they will not ask more than what is compatible with others' interests, and they will not give more than what is compatible with their own interests. In line with this prediction, laboratory experiments suggest that people think that others have a duty to help when mutual aid costs very little and can save lives, but not when the costs are high compared to the benefits (Baron & Miller, 2000). In the same way, individuals need to adjust their help to the number of members. If the group is small, they can give a lot of help to each member whereas, in larger groups, they need to divide this help in smaller shares and give less to each member. Conversely, in small group, individuals can ask more to others than in bigger groups where they are lots of people to help. In line with this idea, observations about rescue in war (Varese & Yaish, 2000) or natural disasters (Singer, 1972; Unger, 1996) suggest that people feel that they have more duty to help when the same situations is framed as involving a small group (typically the helper and the person to help) than when it is framed as involving larger group (a group of helpers and a group of persons to help). This effect may also explain why

people feel they have more duty toward their friends than toward their colleagues, toward their colleagues than toward their fellow citizens, and so on (Haidt & Baron, 1996). To sum up, the contractualist theory explains why individuals do not ask more than what is compatible with others' interests, and do not give more than what is compatible with their own interests.

2.3 Fairness and punishment

The contractualist theory explains a range of cooperative behaviours, people's criteria for giving to others, and how much they will give. What does it say about punishment? First, as we have seen, it contends that cooperation is enforced not by punishment but by the need to attract potential partners. Thus, uncooperative action should mostly trigger partner switches rather than punishment. I have argued elsewhere (Baumard, 2010a) that it is indeed the case. In small scale societies wrongdoers are not punished by altruistic individuals willing to defend the interests of the group. When someone harms or steals someone else, most of the time nothing is done to punish the culprit. If the wrongdoing is very serious and threatens the safety of the victim, she may retaliate in order to preserve her reputation or deter future aggression. In most cases, however, people simply stop wasting their time interacting with immoral individuals. In large scale societies, punishment is much more central. However, institutional punishments do not require any altruism from the members of institutions as they are organized in such a way that members of institutions have a personal interest in punishing (Ostrom, 1990). Thus, although large scale institutions do rely on punishments, they do not rely on an instinct for punishment.

It does not follow from the fact that cooperation has not evolved by punishment that people should never punish wrongdoers. Indeed, we have some duties toward others. As we have seen in section 2.2, we have the duty to rescue others when they are in danger (otherwise we can be liable for what happens to the person in danger) and we have a duty to prevent crimes (otherwise, we can be considered as accomplices). In the same way, we have a duty to restore fairness, otherwise we can be considered to be complicit in the unfair situation. This is the reason why people vote and protest to punish financial firms guilty of misbehaviours or why they give money to NGO such as Amnesty International or Oxfam. That is the best they can do in a society where there is no institutions to easily punish criminals.

Although this duty to restore fairness may explain why people punish wrongdoers, it is quite limited and may not play an important role in the evolution of cooperation. Indeed, we not have to sacrifice ourselves to protect the rights of others, just as our duty to rescue others

or to prevent crimes is limited (we are not obliged to risk our lives to save strangers or to be killed to prevent a robbery—this would be great, even heroic, but surpasses the requirements of duty). Most people consider that it is good but beyond duty to risk one's life to support political dissidents in foreign countries for example. Since punishing others is often very costly (except in economic games where it costs only a couple of euros and where there is no threat of retaliation), we should not observe many of the costly punishments examined by evolutionary scientists (e.g. physically attacking the wrongdoer). The only cases where costly punishments are expected to be common are when punishment coincides with retaliation (i.e. the victim is also the punisher and has a direct interest in defending her life and her possessions) or when some institution supply the cost of punishment by rewarding the punishers (through privileges, gifts, wages, etc.).

The contractualist approach makes a specific prediction about punishment: when it is carried out (mostly by retaliating individuals or by the penal system, but also by the motivation to be fair), it should respect the logic of fairness. When an individual harms or steals from another individual, he is unfair to him. He takes more than he deserves. The respect of fairness implies that the wrongdoer should compensate the victim or he should incur a cost proportionate to the cost he inflicted to his victim. Here, punishment or restorative justice is just the symmetric counterpart of sharing or distributive justice. In the same way that people give resources to others when others are entitled to these resources (e.g. because others have contributed to the production of the resources or because they had previously given some resources), people take some resources from others because others were not entitled to these resources (e.g. because they had stolen them or taken more than what they had contributed to produce). A wrongdoing creates an unfair relationship between the culprit and her victim, and people have the intuition that something should be done to restore the balance of interests—either by harming the culprit or by compensating the victim. Justice should be restored.

To sum up, group selection theory and partner choice theory have two different views on punishment. In the following sections, we will review three kinds of studies of punishment: behavioural games, ethnographic observations and cognitive experiments. In each case, we will compare each theory's predictions and decide to what extent they fit with empirical data.

3. Punishments in behavioural games

3.1 Punishment in public good games

In economic games, punishment has been mainly studied with the public good game (PGG). A typical public good game consists of a number of rounds, say 10. The participants are told about the total number of rounds, as well as all other aspects of the game. The participants are paid their winnings in real money at the end of the session. In each round, each participant is grouped with several other subjects—say three others—under conditions of strict anonymity. Each participant is then given a certain number of "points," say 20, redeemable at the end of the experimental session for real money. Each participant secretly chooses how many of its private tokens to put into the public pot. Each participant keeps the tokens she does not contribute plus an even split of the tokens in the pot (experimenters multiply the number of tokens in the pot before it is distributed to encourage contribution)

A self-interested player will contribute nothing to the common account and will still benefit from the public good. However, only a fraction of subjects conform to the self-interest model. Subjects begin by contributing on average about half of their endowment to the public account. The level of contributions decays over the course of the 10 rounds, until in the final rounds most players are behaving in a self-interested manner (Ledyard, 1994). When the PGG is played repeatedly with the same partners, the level of contribution declines towards zero, culminating in most subjects refusing to contribute to the common pool (Andreoni, 1995; Fehr & Gächter, 2002).

Further experiments have shown that participants are ready to punish others (to fine them) at a cost to themselves (Fehr & Gächter, 2002; Yamagishi, 1986). When costly punishment is permitted, cooperation does not deteriorate. When punishment is not permitted, however, the same subjects experience the deterioration of cooperation found in previous public goods games.

Can participants' punishments allow us to disentangle the contractualist and the group selection accounts of punishment? The two theories have very different views on punishment. According to the group selection theory, punishment aims at sustaining cooperation and increasing the group's welfare (Boyd, Gintis, Bowles, & Richerson, 2003; Fehr & Gächter, 2002; Henrich & Boyd, 2001). On the contrary, in the contractualist theory, there is no need for punishment to be used to sustain cooperation, because competition among partners is supposed to be strong enough to select for cooperative tendencies. When individuals are not

cooperative, they are simply abandoned by others who prefer to join more advantageous collaborators.

This does not mean that people should never punish wrongdoers. As we saw in section 2, we have a duty to restore fairness, otherwise we may become accomplices of the unfair situation. Remember, however, that this duty is limited. This explains why very costly punishments are rarely observed, except if further circumstances (revenge, penal system) combine to bring about punishment. In economic games, however, punishments are not very costly (a couple of euros) and we may observe punishment motivated by pure fairness.

Thus, the contractualist theory predicts the existence of restorative punishment aiming at suppressing unfairness, while group selection predicts altruistic punishments aiming at sustaining cooperation. This difference in the predicted nature and role of punishment allows us to tease apart the validity of the two theories. If the contractualist theory is right, we should observe the same level of punishment whether or not a public good is at stake. On the contrary, the group selection theory predicts that participants should punish only when the production of the common good is threatened.

Dawes et al. (2007) use a simple experimental design to examine whether individuals reduce or augment others' incomes when there is no cooperation to sustain. They call these behaviours 'taking' and 'giving' instead of 'punishment' and 'reward' to indicate that income alteration cannot change the behaviour of the target. Subjects are divided into groups having four anonymous members each. Each player receives a sum of money randomly generated by a computer; the distribution is thus arbitrary (and therefore unfair since lucky players do not deserve a bigger amount of money than unlucky players). Subjects are shown the payoffs of other group members for that round and are then provided an opportunity to give 'negative' or 'positive' tokens to other players. Each negative token reduces the purchaser's payoff by one monetary unit (MU) and decreases the payoff of a targeted individual by three MUs; positive tokens decrease the purchaser's payoff by one MU and increase the targeted individual's payoff by three MUs. Groups are randomized after each round to prevent reputation from influencing decisions; interactions between players are strictly anonymous and subjects know this.

Their results show that individuals incur costs reducing and augmenting others' incomes despite the fact that round after round, participants can see that their behaviour has no effect on the subsequent distribution. Analyses show that participants were mainly

motivated by fairness motives, trying to achieve an equal division of wealth[1]. 68% reduced another player's income at least once, 28% did so five times or more, and 6% did so ten times or more. Also, 74% of participants increased another player's income at least once, 33% did so five times or more, and 10% did so ten times or more.

Most (71%) negative tokens were given to above-average earners in each group, whereas most (62%) positive tokens were targeted at below-average earners in each group. More precisely, subjects who earned ten MUs more than the group average received a mean of 8.9 negative tokens compared to 1.6 for those who earned at least ten MUs less than the group. In contrast, individuals who earned considerably less than other group members received sizeable gifts. Subjects who earned ten MUs more than the group average received a mean of 4 positive tokens compared to 11.1 for those who earned at least ten MUs less than the group. Overall, the distribution of punishment displays the logic of fairness: The more money a participant receives, the more others will "tax" her. Conversely the less she receives, the more she gets "compensated."

Finally, in an additional experiment, the authors presented subjects with hypothetical scenarios in which they encountered group members who obtained higher payoffs than they did. Subjects were asked to indicate on a seven-point scale whether they felt annoyed or angry (1, 'not at all'; 7, 'very') by the other individual. In the 'high inequality' scenario, subjects were told they encountered an individual whose payoff was considerably greater than their own. This scenario generated much annoyance: 75% of the subjects claimed to be at least somewhat annoyed, and 41% indicated that they would be angry. In the 'low-inequality' scenario, differences between subjects' incomes were smaller, and there was significantly less anger. Only 46% indicated they were annoyed and 27% indicated they were angry. Individuals apparently feel negative emotions towards high earners, and the intensity of these emotions increases with income inequality. Moreover, these emotions seem to influence behaviour. Subjects who said they were at least somewhat annoyed or angry at the top earner in the high-inequality scenario spent 26% more to reduce above-average earners' incomes than subjects who said they were not annoyed or angry. These subjects also spent 70% more to increase below-average earners' incomes.

In a subsequent experiment, the same team examined the relation between the random inequality game and the PGG (Johnson, Dawes, Fowler, McElreath, & Smirnov, 2009). They

[1] To be sure that reciprocation was not a motivation, they conducted additional analyses. Results show that negative tokens sent were not significantly affected by negative tokens received in the previous round, and positive tokens sent were not significantly affected by positive tokens received in the previous round.

used an experimental design in which subjects play two games: A random income game measuring inequality aversion, and a modified public goods game with punishment. Their results suggest that those who exhibit stronger preferences for equality are more willing to punish free-riders in PGG. The same subjects who assign negative tokens to high earners in the random income experiment also spend significantly more on punishment of low contributors in the PGG[2], suggesting that punishment in PGG is about not only sustaining cooperation, but also inequality.

In a second condition, participants also had the opportunity to pay in order to help others, and the results were nearly identical. The participants who engage in costly giving to low earners (earning less than the group average) in the random income game send more punishment to the lowest contributor in a group in the public goods game with random payoff. Moreover, those who reduce the income of above average earners and those who increase the income of below average earners in the random income game are both more likely to punish low contributors in the normal PGG.

In another experiment, Leibbrandt and López-Pérez (2008) have studied which distributions are punished in a dictator game. They results show that in many cases, participants' punishment fit better with the contractualist theory than with the utilitarian theory. For instance, many participants think that the equal distribution (150/150) is OK while the unequal distribution (590/60) although much better from a utilitarian point of view is not acceptable and should be punished. Strikingly, many participants are also willing to punish the recipient if he is the one who gets the bigger part of the unfair distribution! This clearly show that participants do not expect others do sacrifice to maximise the welfare of the group. On the contrary, think that the money should be equally distributed.

3.2 Anti-social punishments in public goods games

Anti-social punishment has been observed in many studies, and remains counter-intuitive: Why do some participants punish those who give more than others to the common pool? In a recent study, Herrmann et al. (2008) ran a PGG with punishment in 16 comparable participant pools around the world. They observed huge cross-societal variations. Some participant pools punished the high contributors as much as they punished the low

[2] To verify that envy was not a motivation, they compared the willingness to punish high earners in the random game when high earners are above the participants' income (envy) than above the group's average income (fairness). Analyses show that fairness does a much better job predicting punishment in the public goods game. In particular, when the participant's own income is taken as a reference point, the relation between the willingness to punish high earners in the random game and the willingness to punish high earners in the public good games ceases to be significant.

contributors, whereas in others, people only punished low contributors. In some participant pools, antisocial punishment was strong enough to remove the cooperation-enhancing effect of punishment. Such behaviour completely contradicts the logic of group selection. If a participant is self-interested, he should neither contribute, nor punish. If he is altruistic, he should contribute and punish those who do not contribute or contribute less than average. By contrast, fairness offers a possible explanation for anti-social punishment.

We have seen that, according to the contractualist theory, people punish to restore justice. Therefore, it may be the case that participants who punish high contributors do so because they want to correct a previous unfair action. In PGG, participants have to decide the amount of money they want to give to the common pool. In other words, they have to decide their level of cooperation. Participants may disagree about this level. Some may think that you should give half of your money, others only a quarter, etc. Imagine that you think that a quarter is fine and that you end up being punished by someone who contributed more and thinks that you should have contributed more as well. In this case, you may think that you did not deserve this punishment and that justice should be restored by punishing those who have unjustly punished you.

Herrmann et al.'s extensive study supports this interpretation. First, they observe that it is where contributions are low that participants punish high contributors: The lower the mean contribution in a pool, the higher the level of antisocial punishment. This observation supports the idea that participants punish high contributors because they think the contribution should be lower than the high contributors' level of cooperation. Second, in line with this an interpretation, they observe that participants who punish high contributors are the ones who had low norms of cooperation at the beginning (i.e., participants who had given small amounts in the first rounds). Third, this theory proposes that people punish high contributors because they think they have been unjustly punished by other participants for their low contribution (in their mind, this low contribution is morally acceptable). And indeed, Herrmann et al. find that antisocial punishment increases as a function of the amount of punishment received. Finally, anti-social punishers do not increase their own level of contribution when they are punished for a low contribution, which suggests that they think that it was illegitimate to punish them for their low contribution. All these observations converge to the interpretation of anti-social punishment in terms of fairness. Participants punish high contributors because they feel they have been wrongly punished for their low contribution and that justice should be restored.

Finally, Herrmann et al. found that norms of civic cooperation are negatively correlated with antisocial punishment. More specifically, they constructed an index of civic cooperation from data taken from the World Values Survey. The index is derived from answers to questions on how justified people think tax evasion, benefit fraud, or dodging fares on public transport are. The more worthy of reproach these behaviours are in the eyes of the average citizen, the higher is the society's position in the index of civic cooperation. What they found is that antisocial punishment is harsher in societies with weak norms of civic cooperation. In these societies, people felt unfairly treated by high contributors who require too much from others. This observation fits nicely with qualitative research findings. for instance, In a recent article Gambetta and Origgi (2009) described how Italian academics agree to deliver low contributions, and regard high contributors as cheaters who treat others unfairly by requiring too much of them.

This conclusion does not mean that fairness is the sole determinant of punishment in PGG. Participants may punish for multiple motives, including selfish reasons. Indeed, some participants punish both high and low contributors in order to increase their own relative pay-off, thus acting out of "spite" (Cinyabuguma, Page, & Putterman, 2004; Falk, Fehr, & Fischbacher, 2005; Saijo & Nakamura, 1995). Others may punish for strategic reasons, thinking that, in a repeated PGG with only four participants, threatening others by punishing low cooperators in order to create incentives to contribute to the common good is a good investment (but see Falk, et al., 2005). Finally, some may also punish to defend their reputation even if the experiment is anonymous (Kurzban & DeScioli, 2009).

To conclude, the existence of punishment in PGG does not threaten the contractualist theory of cooperation. On the contrary, several experiments support the existence of "retributive punishments." If this conclusion is correct, then it may be the case that punishment is not an adaptation to support cooperation, but rather a consequence of the evolution for fair relationships. In line with this conclusion, experimental studies suggest that punishment does not always sustain cooperation in PGG. For instance, in Herrmann et al.'s study, punishment did not have an equally strong disciplinary effect on free riders in all participant pools, and that in some participant pools, punishment even had no cooperation-enhancing effect at all (Herrmann, et al., 2008).

4. Punishments in societies without penal systems

In economic games, punishment is relatively cheap: a couple of euros and no threat of retaliation or cessation of relationships. This is not the case in societies where there are no

penal systems and where punishment would lead to retaliations. In this case, the contractualist theory suggests that people may not have the duty to risk their life in order to punish wrongdoers. And indeed, as Evans-Pritchard noted, in societies when there is no penal system, “self-help, with some backing of public opinion, is the main sanction” (Evans-Pritchard, 1940). People have the duty to support punishment. They do not have the duty punish.

In societies without penal systems, criminals are sometimes punished however because the victim and her family has an interest in retaliating in order to defend her life and her possession and to get the reputation of someone who does not tolerate aggressions. This is what Wiessner observed among the Ju/'hoansi: “all people as autonomous individuals are expected to stand up for their rights in dyadic relationships” (Wiessner, 2005, p. 135)

Although punishments in societies without penal systems can be explained by self-serving motives, they differ from the kind of retaliation observed among animals. Indeed, animal retaliations do not have to be fair. In animal societies, an individual can kill another if killing is the best option to deter future attacks. In human societies, however, retaliations have to respect certain limits. Indeed, social interactions are regulated by the logic of fairness. This is the logic of the famous *Lex Talionis* “an eye for an eye, a tooth for a tooth”. Although this rule sounds almost amoral today, it was seen as progress in Antiquity, since it departs from the amoral logic of retaliation. I thus propose to distinguish *retaliation* from *revenge*. Both behaviours are carried out by individuals defending their reputation, but although the first only aims to deter future attacks, the second also has to respect the logic of fairness. Revenge can only take place when one of the parties has been unfairly attacked, and it can only inflict costs proportionate to the prejudice (see for instance Miller, 1990). Note that retaliation and revenge coexist in human societies. Retaliation is common in warfare or among strangers. By contrast, revenge occurs mainly among people belonging to the same society.

In a contractualist framework, costly punishments are thus essentially driven by self-serving motives. However, they are regulated by the logic of fairness. What is this logic? As we saw in section 2, when an individual steals or harms another individual, he is unfair to him. He takes more than he deserves. The respect of fairness implies that the victim should be compensated or the wrongdoer should incur a cost proportionate to the cost he inflicted to the victim. No more, but also no less. The logic of punishment is thus ‘restorative’. It aims to restore fairness. By contrast, group selection predicts that punishment should aim at sustaining cooperation and deterring crimes.

Data from legal anthropology seems to favour the former theory. Indeed, many writers have discussed the process of law in stateless societies with such expression as “restoring the social balance” (Hoebel, 1954). In one of the first ethnographies on law and punishment, *Manual of Nuer Law*, Howell constantly emphasizes that the purpose of the payment is to ‘restore the equilibrium’ as he puts it, between the group of killer and killed. The logic of equilibrium is apparent in the rare detailed cases of punishment among hunter-gatherers. For instance, Hoebel analyzes a case of punishment reported by Kroeber among the Yurok in the late nineteen century. As revealed by this case, punishments aim at compensating the victim, and the compensation has to be proportionate to the harm done to the victim.

> “In this instance, the family of which M was headman, while not owning the beach, possessed long established rights that the flippers of all sea lions caught along the Pacific coast for a distance of about four miles in either direction from its settlement be yielded to it. A hunter named L, disregarded his duty in this respect on several occasions. M, instead of taking legal steps, brooded, and finally assaulted the father of L, wounding him with an arrow. The family of L took action for assault damages. Crossers handled the case in the regular manner. Their verdict was that the damages sustained by virtue of the wounding were slightly less than the damages arising from the violation of M family’s rights for the sea lion flippers. Ergo, the L family’s claim was nullified. The affair was thus presumed to be equitably settled; but, though the claims were adjudicated, the sense of grievance was not washed out. So L nursed his sense of grievance, and, two days after the legal settlement, he cursed M. To lay a curse without legal justification is a violation of duty in Yurok law. On this foundation M entered a claim for damages against L for violation of his duty to refrain for cursing. Crossers were at work when two hotheaded relatives of M killed L. They were terribly in wrong here, for they simultaneously disregarded due process and applied a penalty that was over strong (...). When the sister of L retaliated by cursing the killers, it was a right to do so. But with greater effect, L’s mother entered a claim to the hereditary right for the sea-lion flippers transferred to her as equivalent to her son’s wergild. She won the award.” (Hoebel, 1954, p. 55)

First, this example clearly illustrates the way punishment works when there is no penal system. The aggrieved party has to defend its right by herself. For the Yurok, M was perfectly in her right by attacking someone who disregarded his right on the flippers of sea lions.

Although some institutions condemn the wrongdoers, third parties do not inflict a cost to the wrongdoers at a cost to themselves. It is to the aggrieved party –who has lot to gain in asserting her rights- to defend herself. Second, this example illustrated the fact that punishment does not aim at deterring cheating but rather at restoring fairness. What is at stake in Yurok's debates is whether the wounding inflicted by M is proportionate to the rights violated by L, whether the killing of L is proportionate to the curse laid by L on M or whether the transfer of the hereditary rights on the sea-lions flippers is a fair compensation for the killing of L. Punishments and counter-punishments all aim at restoring the 'equilibrium'.

Restorative punishment is not specific to hunter-gatherers. As Maine noted in *Ancient Law*, punishment is more organised in tribal societies, it is essentially about compensation. Law in Germanic tribes is a good example (for another example tribal law, see Friedman, 1979 on mediaval Iceland).

> "In general, little distinction was made between suits of a civil or a criminal nature-or it more accurately might be said that all offences were treated as if they were civil offences (...). the barbarian code set out in minute what 'compensation' (or composition) would have to be paid by the offending." (Drew 1973, p. 8 quoted by Black, 2000)

The analysis of punishment of adultery among the Ifugao, a Philippine tribe observed by Barton in the early twentieth century, makes the restorative logic transparent. Indeed, in case of adultery, the adulterer has to pay damages in two ways:

> "to the in-laws of his partner in adultery and also to his own wife's kinsmen as a penalty for the breach of his own marital contract. The same holds for a married adulteress as well. Adultery is, as noted, a ground for divorce, but it not be so used. However, if the marriage is to be continued, the offender must then put up a 'general welfare' feast at which he regales both his wife's and his own kinsmen. Eating together restores and renews the equable relations of the two groups." (Hoebel, 1954, p. 119)

The same logic holds for rape.

> "Rape of a married woman by a married man offends both her own and her husband's kin group. Each collect damages equivalent to those paid in a case of aggravated adultery. And then, if the rapist is married, he pays not only to the woman's, her husband's, but also his wife's kin damages' that go with aggravated adultery." (Hoebel, 1954, p. 120)

This example clearly shows that punishing rape is about compensating the victims. The bigger the prejudice is, the bigger the punishment should be.

Further observations go in the same direction. For instance, anthropologists have noticed that many stateless societies punish unintentional or accidental acts. Among the Nuer for instance, a killing, even unintentional or accidental, has to be compensated (Howell, 1954). From the group selection point of view, this phenomenon, called 'strict liability,' seems puzzling. Indeed, if punishment was meant to deter criminals, it would not target accidental or unintentional killing. As Elster remarks, this is seemingly a waste of resources.

> "One might counter that norms of strict liability have good outcome on the whole, even if on particular occasion they may appear absurd. This is manifestly false, however: no good incentive effects are created by a social norm that makes people responsible for all actions in which they are causally involved. Rather the effect is to make people excessively cautious, to the point of paralyzing any initiative." (Elster, 2007, p. 119)

By contrast, it makes sense from the point of view of fairness. Indeed, as Howell puts it, a killing, even unintentional or accidental, disturbs the balance of blood due between the groups, and this has to be redressed (Howell, 1954). This does not mean that the Nuer did not take intentions into account. Even where liability is absolute, in that compensation must be paid, attention may have to be paid to the mental elements of the killing insofar as this may determine whether the liability is discharged by blood for blood, or by payment of cattle. Intention helps to assess the importance of injustice. Say for instance that X has bought a dangerous animal (a tiger). In one case, he is walking in the street with his tiger. And while Y is passing in the street, he is killed by the tiger. In the other case, X locks his tiger in his house, knowing that it may harm someone. Y heard about the tiger and despite warnings, he decide to break in X's house. Eventually, he opens the tiger's cage and is killed. In both cases, X did not want to kill Y. However, in the second case, he took care (and incur a cost) of

protecting people from his tiger. Although he is guilty of killing Y in both cases, he is much less guilty than in the first case in which he did not take care of protecting others.

5. Punishments in societies with penal systems

As we have seen in the previous section, restorative law is thus very primitive. This restorative logic remained central for a long time (for tribal societies, see for instance Gluckman, 1955). In the early Middle Ages, even homicide still only required compensation to the victim's family (Black, 2000). The rise of the power of the State, however, has changed punitive behaviours. In the late Middle Ages, States became able to use force to increase safety, and took the place of the victim in the punishment process. But although the goal of the States is often to deter crimes, restorative logic may not have disappeared. The contractualist theory suggests that although modern societies punish wrongdoing for security and safety reasons, judges and policemen, as well as electors, may still favour a restorative justice over a utilitarian one and may have modified penal system in a contractualist way. Is it really the case?

In this section, we shall examine in detail how people view punishment in modern society. We will follow the same line of inquiry as for behavioural experiments and compare the prediction made by contractualist and group selection theories. Group selection theory predicts that punishment aims to promote the welfare of the group (Boyd, et al., 2003; Fehr & Gächter, 2002; Henrich & Boyd, 2001). This implies that punishment should be calibrated to help the group and deter crime. Here, group selection parallels the utilitarian doctrine of punishment, which contends that punishment should be used to deter crimes and maximise the welfare of the society (Polinsky & Shavell, 2000; Posner, 1983). The utilitarian theory of punishment considers, for instance, that the detection rate of a given crime and the publicity associated with a given conviction are relevant factors in assigning punishments. If a crime is difficult to detect, the punishment for that crime ought to be made more severe in order to counterbalance the temptation created by the low risk of getting caught. Likewise, if a conviction is likely to get a lot of publicity, a law enforcement system interested in deterrence should take advantage of this circumstance by "making an example" of the convict with a particularly severe punishment, thus getting a maximum of deterrence for its punishment. By contrast, the contractualist theory predicts restorative punishment. A crime creates an unfair relationship between the criminal and her victim, so that people have the intuition that something should be done to restore the balance of interests—either by harming the criminal

or by compensating the victim. In intuitive terms, someone is punished because she "deserves" to be punished.

Are the data consistent with one of these theories? Recent empirical studies, relying on a variety of methodologies, suggest that when people punish harmdoers, they generally respond to factors relevant to a retributive theory of punishment (magnitude of harm, moral intentions) and ignore factors relevant to the group selection theory (likelihood of detection, publicity, likelihood of repeat offending) (Baron, Gowda, & Kunreuther, 1993; Baron & Ritov, 2008; Carlsmith, Darley, & Robinson, 2002; Darley, Carlsmith, & Robinson, 2000; Glaeser & Sacerdote, 2000; McFatter, 1982; J. Roberts & Gebotys, 1989; Sunstein, Schkade, & Kahneman, 2000). Darley et al. (2000), for instance, examined intuitions regarding the punishment of prototypical wrongs (e.g., violence, murder, etc.) and found that participants adjusted their sentences in response to changes in the moral status of the offender, the magnitude of the harm, and the reasons the perpetrator committed the harm in the first place (for instance, if she committed the crime in order to help someone or for her own benefit). On the contrary, they generally ignored information about whether she was likely to commit them again in the future.

A subsequent study confirmed these results by looking at the kind of information participants want to acquire (Carlsmith, 2006). For example, when they chose to acquire information about the embezzler's motive, they could learn that the embezzlement had been done to get funds either to continue to lead a dissolute life or to redistribute it to the poor. After they had acquired each piece of new information, participants rated their tentative sentence and their confidence in their sentence. The results show that people tended to first seek out information about just desserts, and then to later seek out incapacitative information (information related to preventing the criminal to commit another offense). In any case, information relevant to deterrence was rarely examined. Sequential judgments of confidence were also affected more by the just deserts information, and less so by the incapacitation information.

A similar result emerged from a test that asked participants to assess penalties and compensation separately for victims of birth-control pills and vaccines in cases involving no clear negligence (Baron & Ritov, 1993). In one set of cases, a corporation that manufactures vaccines is being sued because a child died as a result of taking one of its flu vaccines. Participants were given multiple versions of this case. In one version, participants read that a fine would have a positive deterrent effect and make the company produce a safer vaccine. In a different version, participants read that a fine would have a "perverse" effect, causing the

company to stop making this kind of vaccine altogether (which is a bad outcome given that the vaccine in question does more good than harm and that no other firm is capable of making such a vaccine). Participants indicated whether they thought a punitive fine was appropriate in either of these cases and whether the fine should differ between those two cases. A majority of participants said that the fine should not differ at all, which suggests that they care less about the *effect* of the fine than about the very fact that the corporation has to pay for its fault. In another test of the same principle, participants assigned penalties to the company even when the penalty was secret, the company was insured, and the company was going out of business, so that (participants were told) the amount of the penalty would have no effect on anyone's future behaviour (Baron, et al., 1993; Baron & Ritov, 1993). In all these studies, most participants, including a group of judges, "did not seem to notice the incentive issue" (Baron, 1993, p. 124).

Finally, Sunstein et al. (2000) assessed whether people want optimal deterrence. In the first experiment, participants were given cases of wrongdoing, arguably calling for punitive damages, and also were provided with explicit information about the probability of detection. Different participants saw the same case, with only one difference: the probability of detection was substantially varied. The goal was to see if participants would impose higher punishments when the probability of detection was low. In the second experiment, participants were asked to evaluate judicial and executive decisions made to reduce penalties when the probability of detection was high, and to increase penalties when the probability of detection was low. The first experiment found that varying the probability of detection had no effect on punitive awards. Even when people's attention was explicitly directed to the probability of detection, they were indifferent to it. The second experiment found that strong majorities of respondents rejected judicial decisions to reduce penalties because of a high probability of detection – and also rejected executive decisions to increase penalties because of a low probability of detection. In other words, people did not approve of an approach to punishment that would make the level of punishment vary with the probability of detection. What apparently concerned them was the extent of the wrongdoing, a parameter which is much more relevant in the contractualist theory. Strikingly, these intuitions resist cultural transmission. Respondents at the University of Chicago Law School, who were taught the deterrent theory of punitive awards, still rejected that theory because they thought that utilitarian policies would be unfair.

Using a different methodology, surveys on the death penalty reveal the same pattern. Although many people say that their opinion about the death penalty is based on its

effectiveness as a deterrent (for partisans, it deters crimes, for opponent, it has no effect), several studies have shown that, actually, many people would continue to support the death penalty even if it had no deterrent value (Ellsworth & Ross, 1983; Tyler & Weber, 1982). Ellsworth and Ross (1983), for example, found that 66% of those who said they supported the death penalty indicated that they would still support it if it had no deterrent value. These results suggest that deterrence is not the major source of support for the death penalty (for similar result on civil commitment for sexually violent criminals, see Carlsmith, Monahan, & Evans, 2008). People support the death penalty because it seems to them that it is the only proportionate penalty for certain crimes (murder, rape, etc.).

Another phenomenon appears to be in favour of the contractualist theory. That is, people seem to want to make injurers undo the harm they did, even when some other penalty would benefit others more. Baron and Ritov (1993) found that both compensation and penalties tend to be greater when the pharmaceutical company pays the victim directly than when penalties are paid to the government and compensation to the victim is then paid by the government. Baron et al. (1993) found that participants preferred to have companies clean up their own waste, even if the waste threatened no-one, rather than spend the same amount of money cleaning up the much more dangerous waste of a defunct company. Such a phenomenon does not make sense in an group selection framework. Any penalty should be fine as long as it deters future crime. By contrast, in the contractualist theory, people want to restore a fair situation between the criminal and the victim. They are interested in the advantage the criminal has taken over the victim, and they want it to be compensated. Punishment is about fairness, not about deterring crime.

To sum up, laboratory experiments show that people do not base their punitive judgements on deterrence or safety, but rather on fairness. This conflict between public safety and fairness was well analysed two hundred years ago by Adam Smith who took the example of the sentinel who fell asleep while on watch and was executed because such carelessness might endanger the whole army.

> “When the preservation of an individual is inconsistent with the safety of a multitude, nothing can be more just than that the many should be preferred to the one. Yet this punishment, how necessary so ever, always appears to be excessively severe. The natural atrocity of the crime seems to be so little, and the punishment so great, that it is with great difficulty that our heart can reconcile itself to it. And our reaction in this kind of case is to be contrasted with our reaction to the punishment of ‘an ungrateful

> murderer or parricide, where we applaud the punishment with ardour and would be enraged and disappointed if the murderer escaped punishment. These very different reactions demonstrate that our approval of punishment in the one case and in the other are founded on very different principles." (Smith, 1759)

Rationally, we may think that everyone will be better off if we adopted rules such as the death penalty for sentinels who fall asleep while on watch. However, intuitively, we have the feeling that this punishment is too harsh compared to the misdeed (because in Smith's example, the misdeed did not lead to any problem: our feeling would be different if the misdeed had led to the killing of many innocents).

Friedman (1979) provides another example of the counter-intuitive nature of consequentialist punishment. He notes that, rationally, we should favour the death penalty over imprisonment.

> "Replacing a criminal punishment with another that both is more severe and has lower ratio of punishment cost to amount of punishment, while reducing the probability of conviction to maintain the same level of deterrence, lowers both punishment cost and enforcement cost. Hence imprisonment is always dominated by execution and both are dominated by fines and other alternatives." (Friedman, 1979)

Although executions or fines are more efficient, people often feel that they are either too harsh or too light for some crimes.

People's intuitions may explain the evolution of criminal justice from an utilitarian logic imposed by the State to a restorative logic imposed by people. Take for instance the English Bloody Code, a term used to refer to the system of laws and punishments in England from ~1400–~1850 (Archives and Special Collections, 2010; Potter, 1993). Since there was no police force at that time, many crimes carried the punishment of execution. It was thought that people might not commit crimes if they knew that they could be sentenced to death. This was also the reason why executions were public spectacles until the 1860s. The authorities believed that hanging criminals in public would frighten people into obeying the law and refrain from committing crime. Crimes that were punishable by execution at this time included stealing anything worth more than 5 shillings, stealing horses or sheep, or writing a threatening letter. However, despite there being so many crimes punishable by death, it has been estimated that fewer people were hanged in the 18th century than the century before.

Judges and juries thought that punishments were too harsh for many of the criminals, so they became less inclined to find them guilty in court. Judges would frequently under-value stolen goods so that the accused would no longer face the death penalty. Since the law makers still wanted punishments to scare potential criminals, but needed them to become less harsh, transportation and prison became the more common punishment.

Further observations suggest that people's views on punishment are driven by fairness. For instance, during trials, judge and jury do not only care about the kind of crime that is judged. They also care about the defendant's past, what she did in her life, what kind of person she is. As we have seen above, from a consequentialist point of view, people should only care about parameters such as detection rate, etc. On the contrary, the judge and the jury try to evaluate the person as a whole to find the perfect equilibrium between her sentence and her behaviour. In the same way, we usually consider that when the criminal goes to jail, he somehow "pays his debt". There is a paradox here, for the criminal, far from compensating the victim (or society), costs a lot in terms of public spending. This paradox can only be explained if we consider that people have the intuition that the criminal has taken an unfair advantage over the victim, and that jail makes this advantage disappear by harming him in proportion to his crime.

Many advocates of the group adaptation emphasized that people have strong intuitions about punishment. They note that "people's motives for sentencing criminals are found in the area of 'just desserts?' rather than deterrence" (Fehr & Fischbacher, 2004). They use the data described here to conclude that punishment is an adaptation. But they failed to notice the logic of people's intuitions contradict the view that punishment has evolved to sustain cooperation. If punishment had evolved to deter cheating, it would follow the consequentialist logic. But this is not the case, and the preference for "just desserts" clearly favours the view that punishment is based on fairness.

6. Conclusion

In this paper, we have seen that data from legal anthropology, experimental psychology, and behavioural economics all favour the view that humans punish to restore fairness rather than to sustain cooperation. In other words, our analysis suggests that humans do not have a 'punishment instinct.' They rather have a fairness instinct.

If punishment is mainly motivated by fairness, then why is the theory of altruistic punishment so intuitive? I would like to suggest several explanations. First, if we resist the idea that cooperation has evolved by partner choice, it may be because we all live in large-

scale societies where it seems like we can always easily switch to another group and build a brand new reputation. But this possibility is recent. For a long time, people could not take their car and move to another city, or buy a plane ticket and settle in a new country. Even today, it is costly to leave one's whole social network behind and build a new one, and even today it is hard to totally avoid former friends.

The main reason why we resist the idea that punishment is not about sustaining cooperation, I think, is that when we think about cheating and immorality, we automatically think of penal justice and police. This is what people do when they are asked to justify their moral judgements. Although their judgements on penalties are based on fairness, they usually invoke safety and prevention of crime to explain their decisions (Carlsmith, 2008)[3].

But, as we have seen, referring to police and institutions is a poor way to understand our retributive intuitions. During millennia, humans did not have a penal system to protect them, and they only cooperated because doing so is the best strategy to survive in a cooperative species. And this is still the case. Indeed, although the penal system is important, most of us never end up in front of a court, which means that most of our cooperative behaviours with our spouse, our friends, our colleagues is not regulated by the penal system.

To sum up, it may be the case that when we think about punishment, we usually look in the wrong places. The penal system is recent and sustained by specialised institutions. If we want to understand the dynamic of human cooperation, we should look instead to everyday relationships. If we do so, punishment becomes much less salient and its role much less obvious. Indeed, we rarely punish our friends, our relatives or our colleagues. We only signal our discontent and complain to others (McAdams, 1997). In the worst case scenario, we simply end our relationship, either through a big argument or by gradually interacting less and less with him. And we don't do so to help the group, but rather to preserve our own interest.

[3] This may explain why participants in small scale societies punish cheaters to a lesser extent (Marlowe, 2009; Marlowe, et al., 2008). In societies where there are penal institutions (public or private police, courts, fines, etc.), people are used to be less autonomous and are more accustomed to see others intervene in their lives (exactly in the same way as Europeans accept more third party intervention –in particular from the state– than Americans). Third party punishment would be morally more acceptable in larger societies. One can also speculate that, in larger societies, people think that punishment is useful and therefore have a duty to punish while in smaller societies, where people are not used to see the effect of punishment, people think that punishing others is beyond their duty. Alternatively, it is well known that behavioural games are sensitive to framing effects and demand characteristics (Baumard, et al., 2010; Ensminger, 2002; Heintz, 2005). It could be the case that in larger societies where fines and police play a bigger role than in smaller societies, participants are more prone to interpret the possibility to punish as the right way to participate (to please the experimenter, to make sense of the experiment or to enhance their reputation as the authors suggest).

References

Alexander, R. (1987). *The biology of moral systems*. Hawthorne, N.Y.: A. de Gruyter.

Alvard, M. S. (2004). Good hunters keep smaller shares of larger pies. Comment on To give and to give not: The behavioral ecology of human food transfers by Michael Gurven. *Behavioral and Brain Sciences, 27*, 560-561.

Alvard, M. S., & Nolin, D. (2002). Rousseau's whale hunt? Coordination among big game hunters. *Current Anthropology, 43*, 533.

Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review*, 891-904.

Archives and Special Collections, D. U. L. a. D. W. o. F. C. C. (2010). Crime and Punishment in Durham 1750-1900. Retrieved 26 June 2010

Bailey, R. C. (1991). *The Behavioral Ecology of Efe Pygmy Men in the Ituri Forest, Zaire*. Ann Arbor: Museum of Anthropology, University of Michigan.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*, 325-344.

Baron, J. (1993). Heuristics and biases in equity judgments: A utilitarian approach. In A. Mellers & J. Baron (Eds.), *Psychological perspectives on justice: Theory and applications* (pp. 109). New-York: Cambridge University Press.

Baron, J., Gowda, R., & Kunreuther, H. (1993). Attitudes toward managing hazardous waste: What should be cleaned up and who should pay for it? *Risk Analysis, 13*(2), 183-192.

Baron, J., & Miller, J. (2000). Limiting the Scope of Moral Obligations to Help: A Cross-Cultural Investigation. *Journal of Cross-Cultural Psychology, 31*(6), 703.

Baron, J., & Ritov, I. (1993). Intuitions about penalties and compensation in the context of tort law. *Journal of Risk and Uncertainty, 7*(1), 17-33.

Baron, J., & Ritov, I. (2008). The role of probability of detection in judgments of punishment. *unpublised manuscript*.

Barrett, L., Gaynor, D., & Henzi, S. (2002). A dynamic interaction between aggression and grooming reciprocity among female chacma baboons. *Animal Behaviour, 63*(6), 1047-1053.

Baumard, N. (2008). *A Naturalist and Mutualist Theory of Morality.* Ecole Normale Supérieure, Paris.

Baumard, N. (2010a). Has punishment played a role in the evolution of cooperation? A critical review. *Mind and Society*.

Baumard, N. (2010b). *L'origine de la morale Une histoire naturelle du bien et du mal*. Paris: Odile Jacob.

Baumard, N., Boyer, P., & Sperber, D. (2010). Evolution of Fairness: Cultural Variability. *Science, 329*(5990), 388.

Baumard, N., & Sperber, D. (2007). Morale et réputation dans une perspective évolutionniste. *Worshop « Réputation », Fondazione Olivetti, Roma 14 avril 2007*

Black, D. (2000). On the origin of morality. In L. Katz (Ed.), *Evolutionary origins of morality : cross-disciplinary perspectives* (pp. xvi, 352 p.). Thorverton, UK ; Bowling Green, OH: Imprint Academic.

Boehm, C. (2000). Conflict and the Evolution of Social Control. In L. Katz (Ed.), *Evolutionary origins of morality: Cross-disciplinary perspectives*
Thorverton, UK ; Bowling Green, OH: Imprint Academic.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. (2003). The evolution of altruistic punishment. *Proc Natl Acad Sci U S A., 100*(6), 3531-3535.

Brehm, J. W., & Cole, A. H. (1966). Effect of a favor which reduces freedom. *J Pers Soc Psychol, 3*(4), 420-426.
Bull, J., & Rice, W. (1991). Distinguishing mechanisms for the evolution of co-operation. *Journal of Theoretical Biology, 149*(1), 63.
Carlsmith, K. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology, 42*(4), 437-451.
Carlsmith, K. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research, 21*(2), 119-137.
Carlsmith, K., Darley, J., & Robinson, P. (2002). Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology, 83*(2), 284–299.
Carlsmith, K., Monahan, J., & Evans, A. (2008). *The Function of Punishment in the 'Civil' Commitment of Sexually Violent Predators*: SSRN.
Chancellor, R., & Isbell, L. (2009). Female grooming markets in a population of gray-cheeked mangabeys (Lophocebus albigena). *Behavioral Ecology, 20*(1), 79.
Cinyabuguma, M., Page, T., & Putterman, L. (2004). On perverse and second-order punishment in public goods experiments with decentralized sanctioning. *Brown University, Department of Economics-Working Paper*.
Clark, M., & Jordan, S. (2002). Adherence to Communal Norms: What It Means, When It Occurs, and Some Thoughts on How It Develops. *New directions for child and adolescent development, 95*, 3-25.
Clark, M., & Mills, J. (1979). Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology, 37*(1), 12-24.
Darley, J., Carlsmith, K., & Robinson, P. (2000). Incapacitation and Just Deserts as Motives for Punishment. *Law and Human Behavior, 24*(6), 659-683.
Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature, 446*(7137), 794-796.
de Waal, F. (1997). The chimpanzee's service economy: Food for grooming. *Evolution and Human Behavior, 18*(6), 375-386.
Dunbar, R. I. M. (1993). Co-evolution of neocortex size, group size and language in humans. *Behavioral and Brain Sciences, 16*(4), 681-735.
Ellsworth, P. C., & Ross, L. (1983). Public opinion and capital punishment: A close examination of the views of abolitionists and retentionists. *Crime & Delinquency, 29*(1), 116.
Elster, J. (2007). *Explaining social behavior: More nuts and bolts for the social sciences*. Cambridge ; New York: Cambridge University Press.
Ensminger, J. (2002). Experimental Economics: A Powerful New Method for Theory Testing In Anthropology. *Theory in Economic Anthropology*, 59–78.
Evans-Pritchard, E. E. (1940). *The Nuer, a description of the modes of livelihood and political institutions of a Nilotic people*. Oxford,: At the Clarendon press.
Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, 2017-2030.
Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends Cogn Sci, 8*(4), 185-190.
Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137-140.
Fiske, A. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *PSYCHOLOGICAL REVIEW-NEW YORK-, 99*, 689-689.
Frank, R. (1988). *Passions within reason: The strategic role of the emotions* (1st ed.). New York: Norton.

Frey, B. S., & Pommerehne, W. W. (1993). On the fairness of pricing--An empirical survey among the general population. *Journal of Economic Behavior & Organization, 20*(3), 295-307.

Friedman, D. (1979). Private creation and enforcement of law: a historical case. *The Journal of Legal Studies, 8*(2), 399-415.

Gauthier, D. (1986). *Morals by agreement*. Oxford, New York: Clarendon Press ; Oxford University Press.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24*(3), 153-172.

Glaeser, E. L., & Sacerdote, B. (2000). *The Determinants of Punishment: Deterrence, Incapacitation and Vengeance*: SSRN.

Gluckman, M. (1955). *The judicial process among the Barotse of Northern Rhodesia*. Manchester: Manchester University Press on behalf of the Rhodes-Livingstone Institute, Northern Rhodesia.

Gumert, M. (2007a). Grooming and infant handling interchange in Macaca fascicularis: the relationship between infant supply and grooming payment. *International Journal of Primatology, 28*(5), 1059-1074.

Gumert, M. (2007b). Payment for sex in a macaque mating market. *Animal Behaviour, 74*(6), 1655-1667.

Gurven, M. (2004a). Economic games among the Amazonian Tsimane: Exploring the roles of market access, costs of giving, and cooperation on pro-social game behavior. *Experimental Economics, 7*(1), 5-24.

Gurven, M. (2004b). To give and to give not: The behavioral ecology of human food transfers. *Behavioral and Brain Sciences, 27*.

Haidt, J., & Baron, J. (1996). Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology, 26*, 201-218.

Haley, K., & Fessler, D. (2005). Nobody' s watching? Subtle cues affect gener osity in an anonymous economic game. *Evolution and Human Behavior, 26*(3), 245-256.

Hammerstein, P. (2003). Why Is reciprocity So Rare in Social Animals? . In P. Hammerstein (Ed.), *Genetic and cultural evolution of cooperation* (pp. xiv, 485 p.). Cambridge, Mass.: MIT Press in cooperation with Dahlem University Press.

Heintz, C. (2005). The ecological rationality of strategic cognition. *Behavioral and Brain Science, 28*(6), 825-826.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*(208), 79-89.

Henzi, S., & Barrett, L. (2002). Infants as a commodity in a baboon market. *Animal Behaviour, 63*(5), 915-921.

Herrmann, B., Gächter, S., & Thöni, C. (2008). Antisocial Punishment Across Societies. *Science, 319*(1362).

Hill, K., & Kaplan, H. (1999). Life History Traits in Humans: Theory and Empirical Studies. *Annual Review of Anthropology, 28*, 397-430.

Hoebel, E. A. (1954). *The law of primitive man; a study in comparative legal dynamics*. Cambridge: Harvard University Press.

Howell, P. (1954). *A Manual of Nuer Law: Being an Account of Customary Law, Its Evolution and Development in the Courts Established by the Sudan Government*: Published for the International African Institute by the Oxford University Press.

Jensen, K., Hare, B., Call, J., & Tomasello, M. (2006). What's in it for me? Self-regard precludes altruism and spite in chimpanzees. *Proceedings of the Royal Society B: Biological Sciences, 273*(1589), 1013-1021.

Johnson, T., Dawes, C., Fowler, J., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters, 102*(3), 192-194.
Kahneman, D., Knetsch, J., & Thaler, R. (1986). Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *The American Economic Review, 76*(4), 728-741.
Kaplan, H., Hill, K., Lancaster, J., & Hurtado, A. (2000). A theory of human life history evolution: diet, intelligence, and longevity. *Evolutionary Anthropology, 9*, 156 - 185.
Konow, J. (2003). Which Is the Fairest One of All? A Positive Analysis of Justice Theories. *Journal of Economic Literature, XLI* (December), 1188-1239.
Koyama, N., Caws, C., & Aureli, F. (2006). Interchange of grooming and agonistic support in chimpanzees. *International Journal of Primatology, 27*(5), 1293-1309.
Krebs, D. (1970). Altruism: An examination of the concept and a review of the literature. *Psychological Bulletin, 73*(4), 258-302.
Kurzban, R. (2001). The social psychophysics of cooperation: Nonverbal communication in a public goods game. *Journal of Nonverbal Behavior, 25*(4), 241-259.
Kurzban, R., & DeScioli, P. (2009). Adaptationist Punishment in Humans. *Context*.
Ledyard, J. (1994). Public goods: A survey of experimental research. *Public Economics*.
Leibbrandt, A., & López-Pérez, R. (2008). *The envious punisher*: University of Zurich working paper.
Marlowe, F. (2009). Hadza Cooperation. *Human Nature, 20*(4), 417-430.
Marlowe, F., Berbesque, J., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J., et al. (2008). More 'altruistic'punishment in larger societies. *Proceedings of the Royal Society B, 275*(1634), 587.
Marshall, G., Swift, A., Routh, D., & Burgoyne, C. (1999). What Is and What Ought to Be: Popular Beliefs about Distributive Justice in Thirteen Countries. *European Sociological Review, 15*(4), 349-367.
McAdams, R. (1997). The Origin, Development, and Regulation of Norms. *Michigan Law Review, 96*(2), 338-433.
McFatter, R. M. (1982). Purposes of punishment: effects of utilities of criminal sanctions on perceived appropriateness. *The Journal of applied psychology, 67*(3), 255.
Melis, A., Hare, B., & Tomasello, M. (2009). Chimpanzees coordinate in a negotiation game. *Evolution and Human Behavior, 30*(6), 381-392.
Miller, W. (1990). *Bloodtaking and peacemaking : feud, law, and society in Saga Iceland*. Chicago: University of Chicago Press.
Noe, R., & Hammerstein, P. (1994). Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology, 35*(1), 1-11.
Noë, R., van Schaik, C., & Van Hooff, J. (1991). The market effect: an explanation for pay-off asymmetries among collaborating animals. *Ethology, 87*(1-2), 97-118.
Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge ; New York: Cambridge University Press.
Peck, J. (1993). Friendship and the evolution of co-operation. *Journal of theoretical biology, 162*(2), 195.
Polinsky, A. M., & Shavell, S. (2000). The Economic Theory of Public Enforcement of Law. *Journal of Economic Literature, 38*(1), 45-76.
Port, M., Clough, D., & Kappeler, P. (2009). Market effects offset the reciprocation of grooming in free-ranging redfronted lemurs, Eulemur fulvus rufus. *Animal Behaviour, 77*(1), 29-36.
Posner, R. (1983). *The Economics of Justice*. Cambridge: Harvard University Press.
Potter, H. (1993). *Hanging in judgement: religion and the death penalty in England from the bloody code to abolition*: SCM press.

Price, M. E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior, 23*(3), 203-231.

Rawls, J. (1971). *A theory of justice*. Cambridge, Mass.,: Belknap Press of Harvard University Press.

Roberts, G. (1998). Competitive altruism: from reciprocity to the handicap principle. *Proc. R. Soc. Lond. B 265*, 427-431.

Roberts, J., & Gebotys, R. (1989). The purposes of sentencing: Public support for competing aims. *Behavioral Sciences & the Law, 7*(3).

Saijo, T., & Nakamura, H. (1995). The" spite" dilemma in voluntary contribution mechanism experiments. *The Journal of Conflict Resolution, 39*(3), 535-560.

Schopler, J., & Thompson, V. D. (1968). Role of attribution processes in mediating amount of reciprocity for a favor. *Journal of Personality and Social Psychology, 10*(3), 243-250.

Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., et al. (2005). Chimpanzees are indifferent to the welfare of unrelated group members. *Nature, 437*(7063), 1357-1359.

Singer, P. (1972). Famine, Affluence, and Morality. *Philosophy & Public Affairs, 1*(3), 229-243.

Smith, A. (1759). *The theory of moral sentiments*. London,: A. Millar.

Sober, E., & Wilson, D. (1998). *Unto others : the evolution and psychology of unselfish behavior*. Cambridge, Mass.: Harvard University Press.

Sunstein, C., Schkade, D., & Kahneman, D. (2000). Do People Want Optimal Deterrence? *Journal of Legal Studies, 29*(1), 237-253.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Science, 28*(5), 675-691.

Trivers, R. (1971). Evolution of Reciprocal Altruism. *Quarterly Review of Biology, 46*, 35-57.

Tyler, T. R., & Weber, R. (1982). Support for the death penalty; instrumental response to crime, or symbolic attitude? *Law and Society Review*, 21-45.

Unger, P. K. (1996). *Living high and letting die: Our illusion of innocence*. New York: Oxford University Press.

Varese, F., & Yaish, M. (2000). The Importance of Being Asked. The rescues of Jews in Nazi Europe. *Rationality and Society, 12*(3), 307-334.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology, 20*(2), 415.

Wiessner, P. (1996). Leveling the hunter: constraints on the status quest in foraging societies. In P. Wiessner & W. Schiefenhövel (Eds.), *Food and the status quest* (pp. 171-191). Oxford: Berghahn Books.

Wiessner, P. (2005). Norm Enforcement among the Ju/'hoansi Bushmen A Case of Strong Reciprocity? *Human Nature, 16*(2), 115-145.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*(1), 110-116.