Homo homini lupus? Explaining antisocial punishment

Karolina Sylwester[1]

Benedikt Herrmann[2]

Joanna J. Bryson[1]

[1] University of Bath

Department of Computer Science

BA2 7AY Bath

United Kingdom

[2] Behavioural Economics Team

Joint Research Centre

European Commission

ABSTRACT

Punishing group members who parasitize their own group's resources is an almost universal human behavior, as evidenced by multiple cross-cultural and theoretical studies. Recently, researchers in social and behavioral sciences have identified a puzzling phenomenon called "antisocial punishment": some people are willing to pay a cost to "punish" those who act in ways that benefit their shared social group. Interestingly, the expression of antisocial punishment behavior is regionally diverse and linked to the socio-psychological dimensions of local cultural values. In this review, we adopt an ecological perspective to examine why antisocial punishment might be an advantageous strategy for individuals in some socio-economic contexts. Drawing from research in behavioral economics, personality, social psychology and anthropology, we discuss the proximate mechanisms of antisocial punishment operating at an individual level, and their consequences at the group and cultural levels. We also consider the evolutionary dynamics of antisocial punishment investigated with computer simulations. We argue that antisocial punishment is an expression of aggression, and is driven by competition for status. Our review elucidates the possible socio-ecological underpinnings of antisocial punishment, which may have widespread repercussions at a cultural level.

Homo homini lupus? **Explaining antisocial punishment**

"It is not surprising that there should be a struggle in man between his social instincts, with their derived virtues, and his lower, though at the moment, stronger impulses or desires." (Darwin, 1871, p.104)

Recent reports on antisocial punishment have drawn attention to the duality of human nature. *Antisocial punishment* can be defined as paying a cost to reduce the resources of a person whose previous cooperative behavior benefited the punisher and their group. In past research, the focus tended to be on *altruistic punishment* – paying a cost to reduce the resources of a person who previously exploited group resources. Altruistic punishment has become an area of particular interest because it offers a potential resolution of the quest to understand human cooperation. Extensive cooperation in humans, often considered surprising in light of Darwinian natural selection theory, has been investigated in numerous empirical and theoretical studies (e.g. Gintis, Bowles, Boyd, & Fehr, 2005; Henrich et al., 2004). Altruistic punishment, despite its negative proximate motives[1] and, sometimes, detrimental effect on average payoffs,[2] has been proposed as a form of

[1] Rather than turning the other cheek and continuing to cooperate, motivated by anger humans use punishment towards selfish individuals (Fehr & Gächter, 2002).

[2] (Dreber, Rand, Fudenberg, & Nowak, 2008; Wu et al., 2009)

pro-social behavior promoting cooperation (Fehr & Gächter, 2002). Moreover, it inspired a new theory of the evolution of human cooperation - strong reciprocity (Fehr & Fischbacher, 2003; Gintis, 2000). However, more recent investigations of the full range of available and expressed punishment behavior across cultures have highlighted the existence of antisocial punishment. This has led some to reconsider the "dark side" of human behavior, including a tendency for spite and hyper-competitiveness (Abbink & Herrmann, 2011; Abbink & Sadrieh, 2009; Herrmann, Thöni, & Gächter, 2008; Jensen, 2010).

Our review is motivated by the unexplained cultural variation in antisocial punishment revealed by Herrmann et al. (2008). We propose that the high levels of punishment directed toward cooperators in places like Muscat, Athens and Riyadh reflect different pressures in these socio-economic or cultural environments. These pressures affect the perception of group identity, which leads to changes in individual behavior. We argue that, despite lowering absolute levels of resources across a society taken in aggregate, antisocial punishment may constitute a successful individual strategy for establishing social status and receiving its benefits. This ecological interpretation of costly punishment allows us to present it devoid of ethical loading and enables a better understanding of

its functional causes.[3] In addition to proposing and justifying this theoretical framework, we also emphasize some unresolved questions about costly punishment, and offer testable predictions.

The review is organized as follows. We first focus on the various definitions of costly punishment and how they relate to the concept of altruism in different disciplines. Next, we discuss how methodological manipulations of the cost-to-impact ratios of costly punishment affect its use. We observe that the amount of costly punishment meted out to others (in particular, antisocial punishment) is rationally adjusted to exploit its effect of increasing the positive difference between one's own and others' payoffs. In the proceeding sections we discuss antisocial punishment at three levels: cultural, group and individual. At each level, we show how antisocial punishment could bring advantages despite its initial cost. Crucially, the benefits from using antisocial punishment may result from punishers acquiring a higher status within their groups. In the last section, we present the evolutionary perspective on antisocial punishment and its ultimate consequences for a population, as well as, for individuals.

---

[3] Reproductive timing in human females viewed from an ecological perspective is a notable example of how socially undesirable behaviour, such as teenage pregnancies, can be neutrally explained and considered a biologically sensible strategy.

In this review, to fully understand antisocial punishment, we consider both its *proximate* and *ultimate* causes (Scott-Phillips, Dickins, & West, 2011). A *proximate* explanation refers to the mechanism that leads an individual to express a behavior, while an *ultimate* one describes the evolutionary context that resulted in the appearance of (normally, selection for) a behavior or trait. While many authors have shown that this distinction can be difficult to make (Scott-Phillips et al., 2011; Thierry, 2005), drawing it allows us to investigate two complimentary explanations for why antisocial punishment occurs. First, we focus on the workings of antisocial punishment – the proximate mechanisms that drive it; then, we discuss why it might have evolved – the evolutionary dynamics might have caused it. The answer to the former question is provided primarily by experiments using behavioral economics games while the answer to the latter one comes from computer simulations of evolutionary processes.

Costly punishment terminology

Economists, psychologists and biologists often use the same phrases to mean different things. When drawing together knowledge from various disciplines, it is important to precisely determine what is understood by terms such as *altruistic* or *antisocial* punishment in each, and to define the specific usage in the present discussion. Our use of the word *punishment* originates within the context of behavioral economics

experiments, in which researchers typically employ the Public Goods Game (PGG) with punishment, the Ultimatum Game (UG) and/or the Third Party Punishment game (TPP). PGGs can be played one-shot or for multiple rounds (for the implications which follow from this difference, see Hertwig & Ortmann, 2001). They can also be played with or without punishment opportunities. If a sequence of PGGs is played, the player's group membership can be maintained or different participants may be grouped together in each round. In the latter case, any consequences of punishment do not affect the punisher. UGs and TPPs are, typically, only played for a single round.

PGGs represent a social dilemma because the individual's interests are in conflict with the group's interests. In PGGs, a group of individuals can contribute some portion of their allocation to the public pool, which benefits everyone equally. Individuals who do not contribute anything, or contribute less than others, gain a payoff advantage. In PGGs with punishment, after a round of the PGG, individuals can anonymously punish others (usually at a cost-to-impact ratio of 1/3). In UGs, one individual (the proposer) can share an amount of money between themselves and a recipient. After the proposer's offer, the recipient decides whether they accept it, in which case both parties receive the respective amounts. Alternatively, the recipient can reject the offer, in which case no one receives anything. The act of rejection represents the act of costly punishment because both the recipient and the

proposer suffer a cost. TPPs greatly resemble UGs, with the major difference being that the recipient in the TPP is passive and cannot punish. Instead, an extra third person, not benefitting from the split, has an opportunity to spend money on punishing the proposer.

In an experimental setting, people mete out costly punishment towards uncooperative individuals, even when there is no opportunity to interact with them again (Fehr & Gächter, 2002). Such punishment has been dubbed "altruistic" because the punisher decides to pay a fee to reduce the payoff of free-riders, and this action is likely to make free-riders increase their cooperative contributions in future interactions.[4] Hence, in congruence with the biological definition of altruism (West, Griffin, & Gardner, 2007), punishment is costly to the actor and beneficial to the recipient, where the recipients are individuals interacting with the punished person in the future. The biological definition of altruism refers to the lifetime fitness consequences of a behavior, which

[4] Fehr and Gächter's definition of altruistic punishment is encapsulated in the following two quotes: "Punishment may well benefit the future group members of a punished subject, if that subject responds to the punishment by raising investments in the following periods. In this sense, punishment is altruistic." (p.137, Fehr & Gächter, 2002). "Thus, the act of punishment, although costly for the punisher, provides a benefit to other members of the population by inducing potential non-cooperators to increase their investments. For this reason, the act of punishment is an altruistic act." (p.139, Fehr & Gächter, 2002).

are impossible to capture in behavioral economics experiments. For the sake of convenience, we adopt Fehr and Gächter's term "altruistic punishment" to describe a phenomenon occurring in short-term experimental interactions, although we acknowledge that this definition might be misleading (see Sylwester, Mitchell, & Bryson, submitted).

Altruistic punishment requires that (a) punishers suffer a cost for punishing and (b) punished individuals are thereby induced to become more pro-social. Hence, in behavioral economics, the term "altruistic punishment" is defined through the negative economic outcomes to the punisher and positive economic outcomes to the group. When psychological drives are considered, altruistic punishment seems to be motivated not by the altruistic desire to help the group but rather by negative feelings towards cheaters and the willingness to harm them (Fehr & Gächter, 2002). It could be argued that these negative emotions are a consequence of egalitarian preferences and that the underlying psychological motivation is, therefore, altruistic (Cinyabuguma, Page, & Putterman, 2006; Denant-Boemont, Masclet, & Noussair, 2007; Nikiforakis, 2008). However, studies investigating egalitarian preferences typically use games that measure the degree to which people are willing to reduce others' income, rather than their own income. A reduction of others' income is as likely a result of competitive preferences as egalitarian ones. Therefore, it is questionable whether

punishment behavior should ever be considered "altruistic", in the folk-psychological sense.

Researchers working on costly punishment noticed that in behavioral economics experiments some punishment is directed not to free-riders but to cooperators instead (the earliest record of this phenomenon is provided by Ostrom, Walker, & Gardner, 1992). This punishment type has been dubbed, antisocial (Herrmann et al., 2008), spiteful (Falk, Fehr, & Fischbacher, 2005) or perverse (Cinyabuguma et al., 2006). Antisocial punishment, the "sanctioning of people who behave prosocially" (p.1362, Herrmann et al., 2008), is defined in a broader manner than altruistic punishment (see Table 1). Both altruistic and antisocial punishment are costly to the punisher and even more so to the punished, but the definition of antisocial punishment makes no reference to the consequence of such punishment to group cooperation and welfare. Rather, antisocial punishment focuses on the punishment's target: it is the punishment of those who give more than the punisher.

Hermann et al. (2008) found a statistically significant negative correlation between antisocial punishment and cooperative contributions measured across all subject pools. However, as shown in Table 1, antisocial punishment can sometimes be functionally neutral or even altruistic, in the sense that punishing an individual with a higher cooperative contribution can prevent this person from reducing the level

of their contributions or even encourage them to contribute more[5]. Such an effect can be enhanced by the fact that, in PGG, punished individuals typically do not know who punished them. As a result, they may suspect that the punishment came from a cooperator and hence is deserved. This thread of reasoning finds support in Herrmann et al.'s (2008) data. In 12 out of 16 participant pools, receiving antisocial punishment did not correlate negatively with contributions in the following rounds.[6]

In this review, we will stick to the terms "altruistic" and "antisocial" punishment because, although imprecise and ethically loaded, they are well established in the literature. In our opinion, the evidence suggesting the psychologically- or biologically-altruistic character of punishment is weak. In the experimental setting, the altruistic nature of punishment can be identified only when repeated PGGs are played with different participants in each round, or in one-shot TTPs, but even then it is possible to find selfish explanations for punishment, for example spite. Moreover, punishment of free-riders, instead of positively affecting future contributions, may actually decrease

---

[5] Such an effect has been noticed by Herrmann et al. (p.1366, 2008): "Some antisocial punishment can be efficiency-enhancing in intent to induce the punished individual to increase his or her contributions."

[6] See Table S7B in Herrmann et al.'s (2008) supplementary material. Cities where participants decreased cooperation after being a victim of antisocial punishment: Bonn, Minsk, Samara and Istanbul.

them (Sylwester, Mitchell & Bryson, submitted). Therefore, in this review we will use *altruistic* to denote any punishment meted out by cooperators to free-riders. Depending on the study, cooperators are either defined with respect to individual cooperativeness (those who contribute more than, or equally to, another individual are cooperators, while those who contribute less are free-riders) or to average group contributions (those who contribute more than, or equal to, the group mean are cooperators, those who contribute less are free-riders). *Antisocial* will be used as it was defined by Herrmann et al. (2008). Therefore, any punishment imposed by free-riders on cooperators, or individuals of equal contributions, will be referred to as antisocial.

1. The price of punishment

Researchers investigating costly punishment typically assume that punishment is more costly to the punisher than to the punished. Due to convention rather than any particular rationale, the most commonly used cost-to-impact ratio is 1:3; it costs the punisher one point to reduce the payoff of the punished individual by three points. Although costly punishment can be considered irrational from the perspective of maximizing the absolute payoff, it does follow a rational rule when relative payoff is prioritized.

Expenditure on punishment is strongly affected by the cost-to-impact ratio. The general finding is that the use of punishment decreases

as the punishment price increases (e.g. Anderson & Putterman, 2006). Despite this, some costly punishment (mostly directed at uncooperative individuals) is observed even when the cost to the punisher is larger than the cost to the punished individual. Antisocial punishment does occur, though rarely, even with a high relative cost of punishment (Anderson & Putterman, 2006).[7]

There is variation in the results reported concerning sensitivity to the relative cost of punishment. Using data from U.S. participants, Carpenter (2007) analyzed the behavior of free-riders who punished cooperators, cooperators who punished free-riders and free-riders who punished other free-riders.[8] Out of the three groups, free-riders punishing other free-riders were most sensitive to the price of punishment. Free-riders who punished cooperators did not condition their punishment decisions on price. Carpenter's results contrast with those obtained by Falk, Fehr & Fischbacher who used a sample of Swiss participants (2005). These researchers found that when the cost of punishment is the same to the punisher as to the punished, antisocial punishment

---

[7] In Anderson and Putterman's (2006) study there were three price-to-impact conditions with ratios in condition 1: 0/100, 30/100, 60/100, 80/100, 120/100, condition 2: 0/100, 5/100, 10/100, 20/100, 30/100 and condition 3: 30/100, 40/100, 50/100, 60/100, 70/100.

[8] Free-riding was defined as a negative deviation from the group average. Punishment price-to-impact ratios were as follows:1/4, 1/2, 1/1, 2/1, 4/1.

disappears.[9] When punishment resulted in lowering the payoff of the punished person to a greater extent than reducing the cost to the punisher, sanctioning of cooperators by defectors and defectors by other defectors occurred frequently.

Egas and Riedl (2008) varied the cost and the impact of punishment and investigated how such a manipulation affected cooperation and punishment decisions in repeated PGGs played by Dutch speakers from around the world.[10] As in Falk, Fehr and Fischbacher's study, cooperative individuals were willing to punish when the cost to the punisher was equal to, and even when it exceeded the cost to the punished, though in such cases cooperation was not maintained. Unlike in Falk et al.'s study, Egas and Riedl observed antisocial punishment of more cooperative individuals in all cost-to-impact conditions.[11] In agreement with Falk et al.'s results, antisocial punishment was highest when its cost was relatively low in comparison with the impact on the punished (28% of all punishment acts). However, it remained at the level of 22.3% and 18.5% in the two conditions where the cost to the punisher

---

[9] In their study there were two price conditions: a low-sanction condition with a price-to-impact ratio of 1/1 and a high-sanction condition in which the price-to-impact ratio of punishing cooperators was 1/3.33 while punishing defectors was 1/2.5.

[10] The price-to-impact ratios used by Egas and Riedl were: 1/3, 3/1, 1/1 and 3/3

[11] The researchers call this counter-intuitive punishment.

was equal to the impact on the punished. Surprisingly, even when the punishing cost exceeded its impact by three times, antisocial punishment was still present (13% of all punishment acts).

What happens when punishers themselves can decide about the cost-to-impact ratio of their punishment? Theories of inequality aversion (e.g. Fehr & Schmidt, 1999) suggest that the punisher should use a ratio that would result in minimizing the payoff difference between themselves and the punished. However, if punishment is motivated by the desire for revenge, competition or the pursuit of social status, punishers should adjust the ratio in a way to create an inequality favorable to them. A critical test of these predictions was conducted using the Dictator game with punishment, in which recipients were allowed to decide how much money they wished to deduct from the dictator's account and where the cost of punishment to the punisher was always $1. Two-thirds of the resultant punishments were inequality-seeking. That is, the punisher decided to deduct from the Dictator more money than was necessary to maintain equality. One-third did deduct only the amount of money necessary to reach equality or less (Houser & Xiao, 2010).

Researchers have tended to focus on costly punishment where both the punisher and the punished suffer a cost. It is possible to imagine that non-monetary punishment, in the form of a reprimand that does not affect either the punisher's or the punished's payoff, has some effect on cooperation. Indeed, both costly and non-monetary punishment were

found to increase cooperation, but the effect of non-monetary sanctions weakened over time (Masclet, Noussair, Tucker, & Villeval, 2003). As in other studies on costly punishment, monetary sanctioning was predicted by both negative and positive deviation from the punisher's cooperation level, indicating the presence of altruistic and antisocial punishment. However, in the condition where non-monetary sanctions were used, while the effect of altruistic punishment persisted, antisocial punishment was absent. Masclet et al.'s (2003) study is important in that it gives insight into the motivations behind antisocial punishment. The fact that non-monetary reprimands were not used to punish antisocially indicates that the reason for using antisocial punishment is not to change other individuals' future economic behavior but to negatively affect their payoffs.

The presented evidence does not allow for an unequivocal conclusion about how the cost-to-impact ratio of punishment affects antisocial punishment. While some studies show that changing the cost-to-impact ratio affects antisocial punishment to a greater extent than altruistic punishment and that antisocial punishment is more likely to be reduced when the ratio is unfavorable to the punisher, others do not report such an effect. Despite the mixed findings reported in the studies, it appears that antisocial, rather than altruistic, punishment is more sensitive to the manipulations of the cost-to-impact ratio. In line with this conclusion is the fact that sanctioning cooperators does not occur when

their payoffs cannot be altered. Moreover, free-riders who are potential antisocial punishers are less willing to buy costly information about other's contributions than more cooperative individuals who become altruistic punishers (Page, Putterman, & Garcia, 2008). This suggests that some instances of costly punishment, in particular antisocial punishment, may function as aggressive acts, and are not contingent on the previous cooperative behavior of the punished individuals. In sum, in apparently irrational costly antisocial behavior, the decisions to punish are, at least in some studies, logically tied to the effectiveness of such punishment and to the ability to increase the positive difference between others' payoffs and one's own.

Cross-cultural variation in punishment

A human sense of fairness is omnipresent but takes on different forms around the world (Henrich et al., 2005). A cross-cultural analysis of punishment in UGs and of TTP games revealed a consistent trend; as the offered amount approached an equal split, recipients in the UG and observers in TTP were less willing to punish (Henrich et al., 2006). Interestingly, in some societies a small fraction of recipients sanctioned those whose offers were hyper-fair i.e. those who donated more than an equal split would predict. The suggested reason for such behavior, observed mostly in gift-giving cultures, was the reluctance of recipients to feel indebted to the proposers and the subordinate position resulting

from such a debt. In consequence, cooperators were punished antisocially, but, interestingly, in this situation the cost to the punisher was even higher than the cost to the punished.[12]

Punishing generous individuals appeared as a *leitmotiv* in Herrmann et al.'s (2008) cross-cultural study on costly punishment, conducted in 16 comparable subject pools. Participants from different cities across the world played multi-round PGGs, with each round followed by a punishment opportunity. Herrmann et al. (2008) found that the level of antisocial punishment, measured as punishment towards individuals whose PGG contributions were equal to or exceeded the punisher's contributions, varied dramatically across societies. Notably, high levels of antisocial punishment were observed in Greece, Turkey, the former Soviet Union and the Middle East while lower levels were found in the U.S, Australia, the Far East and Northwestern Europe[13]. Previous experiments, conducted in places with low levels of antisocial punishment, showed that the opportunity to punish positively affected group cooperation. However, not surprisingly, in subject pools where

[12] In splits where the proposer offers more than a fair share to the recipient (e.g. 30 for the proposer and 70 for the recipient), a recipient who rejects the offer suffers a higher cost (70) than the "punished" proposer (30).

[13] Scandinavia, the UK, Germany & Switzerland. Southwestern Europe, e.g. France, Spain & Italy were not tested.

cooperators were punished heavily, cooperation levels did not increase with punishment.

In an attempt to explain the observed cross-cultural variation, Herrmann et al. investigated possible relationships between antisocial punishment and a number of socio-demographic factors. Democracy ranking and a measure of the prosperity of a country (GDP per capita) were negatively correlated with antisocial punishment, suggesting that high socio-economic development coincides with the cooperation-enhancing function of punishment. Antisocial punishment was also related to various cultural dimensions of the investigated countries (see Hofstede, 2001) e.g. it occurred more often in places where the inequality in society was high (high Power Distance), where ties between individuals and their in-group are strong (low Individualism), where gender differences tend to fade away (low Masculinity) and where uncertainty avoidance is high.

In their analysis, Herrmann at al. (2008) emphasized two factors as possible explanations for the cross-cultural variation in antisocial punishment: the norms of civic cooperation and the rule of law. The norms of civic cooperation is a measure based on questions used in the World Values Survey describing the strength of abiding cooperative norms in a society and the level of disapproval for breaking them. The rule of law is an indicator developed by the World Bank to describe the extent to which people perceive their government, police, courts and

authorities as fair, trustworthy and effective at law enforcement. Both measures were negatively correlated with antisocial punishment. Additionally, the researchers investigated a link between Inglehart's cultural dimensions "traditional vs. secular-rational values" and "survival vs. self-expression values" and antisocial punishment. They found less antisocial punishment in cities where self-expression values i.e. social liberties and personal freedom mattered more than survival values, which represent economic and physical security.[14]

With so many interdependent predictors of antisocial punishment, it is difficult to determine their relative importance and assess their explanatory power. While Herrmann et al. focused on predictors involving ethical evaluation of certain behaviors by the society (norms of civic cooperation); and the quality, efficiency and fairness of a centralized sanctioning system within a society (rule of law), it is possible to imagine that differences in antisocial punishment are driven by other societal characteristics. For example, if antisocial punishment is proximately motivated by dominance and the desire for social control, it would be reasonable to focus on its relationship with power distance and survival/self-expression values. High levels of antisocial punishment would be expected in places where social hierarchy and demonstration of

[14] This correlation is unsurprising given that Inglehart's "survival vs. self-expression values" are related to Hofstede's power distance and Individualism-Collectivism dimensions (Inglehart & Welzel, 2005).

power play an important role, and in harsher environments where individuals need to focus on local competition with their neighbors in order to succeed.

The variation in cooperation observed in Herrmann et al.'s (2008) data was affected by individual heterogeneity and group-level differences and most importantly by the membership in a "world culture" (Gächter, Herrmann, & Thöni, 2010).[15] Apart from the cultural differences in the average cooperation level when punishment was possible, there were also some interesting differences in the patterns of reacting to punishment. In subject pools with high levels of antisocial punishment, the level of cooperation remained low but relatively stable. In contrast, in places where punishment of free-riders dominated and antisocial punishment was scant, some participants, when the opportunity to punish was introduced, almost immediately increased their pro-social contributions (e.g. Boston, Nottingham, Copenhagen, Bonn, Zurich and St Gallen). In other subject pools the increase in cooperation occurred gradually over the course of rounds (e.g. Seoul, Chengdu and Melbourne). In general, clustering the subject pools according to the Inglehart and Baker (2000) schema did approximate the patterns of the reactions to punishment but

[15] World cultures have been defined following Inglehart and Baker (2000) and Hofstede (2001) as a way to capture their historical and cultural backgrounds.

403 there were exceptions. Melbourne, categorized as an English speaking
404 culture, together with Nottingham, displayed a pattern similar to those
405 observed in the cities of the Confucian culture-type. Boston, on the other
406 hand, resembled the pattern observed in protestant non-English speaking
407 Europe.

408 Running identical experiments with the same experimenter and
409 instructions allows us to unravel cross-cultural variation in antisocial-
410 punishment behavior. By employing a slightly different design, and
411 comparing the behavior in subject pools from two countries, we may
412 illuminate other cross-cultural patterns, not visible using the earlier
413 experimental method. While costly punishment increases cooperation in
414 Boston (Dreber, Rand, Fudenberg, & Nowak, 2008), it does not do so in
415 Beijing (Wu et al., 2009). In contrast, Herrmann et al. (2008) found that
416 the opportunity to use punishment positively affected contributions in
417 both subject pools, and that both Chinese participants from Chengdu and
418 US participants from Boston exhibited similar levels of costly
419 punishment, with only marginally higher level of antisocial punishment
420 in China. Unlike in Hermann et al.'s paradigm with a PGG, in Dreber et
421 al.'s and Wu et al.'s experiments participants had an opportunity to
422 cooperate, defect or punish within a dyad, in each round. Wu et al. (2009)
423 discovered high levels of indiscriminate punishing in China in
424 comparison to the US. The researchers explained the differences between
425 theirs and Herrmann et al.'s study by the differences between protocols

used. In the repeated PGG, Chinese participants might have recognized the concept of reputation, so important in their culture, whereas in the dyadic encounters this concept was not applicable.

Another cross-cultural study, conducted by Gächter and Herrmann (2009), supported their 2008 results. In an experiment comparing antisocial punishment in Swiss and Russian participants, it was confirmed that the punishment directed at cooperators in one-shot games meted out by Russian participants was higher than antisocial punishment in Switzerland.[16] What merits attention is that participants in both investigated regions could accurately predict the levels of antisocial punishment, which suggests that common cultural origins predispose people to correctly assess the cooperative and uncooperative intentions of the members of their cultural group. In Russia, participants exhibited more exploitative behavior in the sense that, even if they expected high levels of cooperation from others, their own cooperative contribution was lower than Swiss participants' contributions. Introducing punishment had a positive effect on cooperation in Switzerland but a detrimental effect on cooperation in Russia. In the latter case, this effect was mostly driven by the change in the behavior of top contributors, who, presumably expecting antisocial punishment, became less cooperative.

[16] The reference level was the group average.

Evidence that an opportunity to punish produces different types of behavior in different cultures is growing. In a recent study, American and Romanian students showed a similar level of cooperative behavior when it was measured by games without punishment (Ellingsen et al., 2012). However, in repeated PGGs with punishment, American students tended to use cooperation-enhancing altruistic punishment, while Romanian students frequently meted out antisocial punishment. Interestingly, Romanian students often used indiscriminate punishment targeting both cooperators and non-cooperators. This finding is in line with our re-analysis of Herrmann et al.'s dataset (Sylwester, Mitchell & Bryson, in preparation), showing a non-exclusive use of antisocial and altruistic punishment.

It is plausible to expect that, within a given culture, socio-demographic factors will modulate the occurrence of antisocial punishment, as they do with cooperation and third-party punishment (Marlowe et al., 2011). In a study conducted in rural and urban Russia, socio-demographic variables were found to affect cooperative but not punishing behavior (Gächter & Herrmann, 2011). High levels of antisocial punishment were unrelated to the age group and region of the sample but, surprisingly, participants with a university degree and those who were members of a voluntary organization exhibited higher levels of

467 antisocial punishment.[17] It is important to note, however, that one-shot
468 games were used in that experiment and different patterns might be
469 revealed if participants are allowed to interact in the same group for a
470 longer period of time, as in Herrmann et al. (2008).

471 So far, the evidence gathered by Herrmann et al. (2008) provides
472 the most complete picture of antisocial punishment in different cultures.
473 The patchwork of other studies that differ in methodology do not
474 facilitate a coherent theory of the driving forces behind the variation in
475 antisocial punishment. The direction of the correlations between
476 antisocial punishment and different socio-economic factors suggests that
477 certain conditions can contribute to its occurrence. More specifically, it
478 appears that antisocial punishment frequently takes place in cultures
479 where the potential cost of it is low in relation to its benefits, for
480 example, in places where norms are frequently infringed, free-riding is
481 commonly approved of and legal sanctioning institutions are not
482 perceived as being fair or efficient. In such places, the potential cost of
483 being caught red-handed when punishing cooperators is low in
484 comparison to places where unethical behavior is strongly penalized and
485 disapproved of by both members of the society and legal institutions. On
486 the other hand, we observe antisocial punishment in places where there is

[17] Though voluntary organisations in the former Soviet Union might have a different character from voluntary organisations in established market economies.

a lot to be gained from acquiring a higher rank in the group (even at a cost of the absolute payoff) and where status and power may have a dramatic impact on the quality of life and survival. In cultures with high power distance the benefits coming from having a dominant status are much higher than where power distance is low. In places abundant in resources and with low inequality, gaining power might bring smaller ecological benefits than in places where resources are low and competition is fierce.

Antisocial punishment at the group level

Variation in antisocial punishment occurs at various levels. Starting from the top, we can consider cultures (e.g. as defined by Inglehart & Baker, 2000), populations within a culture, groups within a population and individuals within a group. Micro-level behavior modulates macro-level, so examining individual drives and social influences within different environments may help explain variation in the cultural make-up. In this section, we discuss between- and within-group competition that may be affecting the observed variation in antisocial punishment. Punishment can be imposed within one's own close social group or it may be inflicted on individuals from another group. Since altruistic punishment enhances group welfare in the long run (Gächter, Renner, & Sefton, 2008) while antisocial punishment can be expected to decrease it, the use of these two types of punishment

towards in-group and out-group members should be contingent on the severity of inter- and intra-group competition.

Inter-group competition

The parochial preferences widely documented in humans manifest themselves in people favoring individuals from their own social group (Tajfel, 1970). In-group favoritism can occur in any situation where an individual has an option to positively or negatively affect another individual's well-being. Hence, we should be able to observe selective use of altruistic and antisocial punishment towards in-group versus out-group members. Costly altruistic punishment might be a useful tool for enhancing a group's cohesion and cooperation, particularly when it is done within one's own social group and not inflicted on out-group members. In contrast, antisocial punishment, which is likely to result in reducing group cooperation and coordination, could be an effective way to gain competitive advantage over another group when inflicted on members of an out-group. This in-group out-group reasoning might be underlying the observed variation in antisocial punishment. Excessive generosity displayed by some individuals can possibly be interpreted as a signal of dominance rather than cooperation. High status of these cooperative individuals distinguishes them form the rest of the group. In consequence, cooperators are not perceived as in-

group members and fall victim of antisocial punishment.[18]

When costly punishment is meted out within one's own group, effective altruistic punishment and inhibited antisocial punishment will positively affect the collective payoffs of individuals as a group This, in turn, can increase the odds of one group gaining advantage over another in between-group competition. Where between-group competition has significant consequences, being a relatively weak individual in a dominant group may be better than being a dominant individual in a subordinate group (Queller, 1994; Wilson, 2004).

The same logic can be applied to a situation when individuals have an opportunity to punish members of an out-group. It is reasonable to expect that with a higher degree of between-group competition the use of antisocial punishment towards out-group members will increase. Directing antisocial punishment to out-group members may undermine the out-group's cooperation or make the mechanism of norm enforcement through altruistic punishment less effective. Either could increase the competitive status of the punisher's own group.

Indeed, experiments conducted in Papua New Guinea with two

---

[18] In a recent study, U.S participants voted to expel from the group not only the most selfish members, but also the ones who excessively contributed to the public good and used little of it (Parks & Stone, 2010). Social comparison mechanisms, combined with the unwillingness to adhere to high norms established by the over-generous individuals, were proposed as explanations for this phenomenon.

548 distinct social groups revealed that altruistic punishment was highest
549 when the person in charge of the split, the recipient and the punisher
550 came from the same social group, and also when only the recipient and
551 the punisher came from the same group (Bernhard, Fischbacher, & Fehr,
552 2006). Most antisocial punishment was observed in the latter case,
553 confirming that punishers were more likely to punish in a way that
554 negatively affected payoffs of an out-group member. In another study
555 with artificially created groups of Japanese participants, a similar pattern
556 was observed (Shinada, Yamagishi, & Ohmura, 2004).[19] Punishing of
557 free-riders by cooperators happened more frequently when done within
558 one's own group (this result was also obtained by McLeish & Oxoby,
559 2007), but, interestingly, free-riders meted out harsher punishment on
560 other free-riders from an out-group rather than in-group. In Shinada et
561 al.'s (2004) study, antisocial punishment was minimal and no in-
562 group/out-group effects were reported. Perhaps this is unsurprising,
563 given Japan's high GDP and the strong rule of law in that country.
564 One-shot TPP experiments have also been conducted in India to
565 investigate the impact of the different caste memberships on punishing
566 behavior. While high-caste participants punished norm violators more
567 severely than low-caste participants (Hoff, Kshetramade, & Fehr, 2009),

[19] The group distinction was created by telling participants that the other members either belonged to their own or a different academic unit.

568 the caste differences in the punishment of cooperators were not
569 significant (Fehr, Hoff, & Kshetramade, 2008). Investigating spiteful
570 behavior of low and high castes, using a series of binary choice Dictator
571 games (in which one person decided about the split of a given amount of
572 money), provided mixed results. When presented with a choice between
573 70/90 (other/self) distribution and 90/90 distribution, 42% of high caste
574 participants and only 21% of low cast participants chose the first
575 (spiteful) option. In contrast, when deciding between 150/150 and
576 100/160 distributions, 83% of high caste and only 53% of low caste
577 participants chose the first (equal split) option (Fehr et al., 2008). In the
578 seven possible choices, high caste participants preferred the spiteful
579 distribution more than low caste participants in only one case (in which
580 the p value was marginally significant). However, the researchers
581 concluded that "high-caste subjects (compared to low-caste subjects) are
582 considerably more likely to reduce others' payoffs if behind, or to take
583 other spiteful actions" (p.499, Fehr et al., 2008).
584 Mere in-group/out-group categorization may not invoke hostility
585 and antisocial sanctions. As argued above, what triggers inter-group
586 conflict and aggression is the social level at which the most significant
587 competition takes place. In a sample from Swiss army platoons, group-
588 membership *per se* did not affect the occurrence of antisocial punishment
589 but resulted in more altruistic punishment when the victim of defection
590 was in-group and the defector was out-group (Goette, Huffman, Meier, &

Sutter, 2010). However, when between-group competition was introduced, costly punishment was mostly imposed on cooperators and free-riders from the out-group. At the same time, in-group cooperation increased. This points to an important role inter-group competition plays in inducing antisocial punishment (and Schadenfreude, see Leach, Spears, Branscombe, & Doosje, 2003). Competition with the out-group can also induce excessive and wasteful punishment of in-group members. Contests between groups resulted in above-rational expenditures on competition but also in high expenditures on within-group punishment of individuals whose financial engagement in the conflict was lower than the group's average (Abbink, Brandts, Herrmann, & Orzen, 2010). High expenditures on costly punishment in the presence of competition have also been found by Sääksvuori et al. (2011).

The levels of antisocial punishment observed in conventional PGG experiments appear to be low when contrasted with the levels towards the out-group members induced by conflict. A possible interpretation of this would be that punishment in ordinary PGG is only a side-effect of mechanisms evolved for conflict situations. The act of costly punishment, when taken out of the PGG context, can be perceived as mere aggression. Engaging in aggressive interactions with out-group members in the presence of conflict may be advantageous, in that it may help preserve a group's resources such as territory. Herrmann et al. (2008) found a negative correlation between antisocial punishment and

scores on the individualism/collectivism dimension. Antisocial punishment occurred more often in places where group identity plays a great role and where, in general, ethnocentrism and xenophobia are more pronounced. A possible explanation for this pattern might be that participants perceived other anonymous players as members of an out-group rather than in-group.

When extracted from the context of PGGs, costly punishment might be an effective weapon used in inter-group conflicts because the cost of aggression is smaller than its consequences to the opponent. Using altruistic punishment in conflict, although still effective at the individual level, might not work in the long term, because it may result in the out-group becoming more cooperative and coherent. Instead, antisocial punishment of out-group cooperators undermines the stability of the other group's social norms.

Intra-group competition

In ecological contexts where intra-group competition is fierce, individuals will use aggression towards members of their own group. Costly punishment typically decreases average payoffs (Dreber et al., 2008; Wu et al., 2009), however, it might be useful for displaying aggression and gaining relative advantage over the punished individual. Previous research has shown that people do care about their relative payoff within a group. For example, Saijo and Nakamura (1995) made participants face a non-dilemma in which the payoff maximizing choice

was to contribute the whole allocation to the public pool[20]. Although the contributions to PGG in the non-dilemma condition were higher than in the standard dilemma, a considerable number of participants still refrained from contributing and failed to maximize their payoff. The average amounts saved in the non-dilemma situation were higher than the average investments to the public pool in the traditional dilemma version of the PGG. This indicates that in the no-dilemma situation more participants chose the non-optimal (non-payoff-maximizing) outcome than in the traditional dilemma, indicating that the non-dilemma may have been taken as a spiteful dilemma.

Saijo and Nakamura (1995) concluded that there exists a population of spiteful individuals who value their ranking within the group more than their absolute payoff. In a similar but more recent study, even when the payoff maximizing decision was to contribute everything to the public pool, a considerable number of participants did not do that (Kümmerli, Burton-Chellew, Ross-Gillespie, & West, 2010). The researchers described this phenomenon as "resistance to extreme strategies" or "imperfections" and discovered that a considerable number of participants perceived their group members as competitors rather than

[20] Saijo and Nakamura (1995) used two variants of the PGG marginal per capita return from each point invested: low return (standard PGG) where each invested point yields 0.7, and high return (anti-dilemma) where each invested point yields 1.43 points.

full collaborators even when competition has been deliberately repressed by the experimental condition. Analogously, the reluctance to accept an unfair split in the UGs described earlier may be dictated not by the norm of fairness but by competitive preferences and/or simple heuristics (Binmore, 2007).

Could this competitive tendency in humans be an artifact of lab experiments using specific homogenous samples (see Henrich, Heine, & Norenzayan, 2010)? Recent studies revealed that "spiteful" punishment (measured as the rate of rejection of offers in the Ultimatum Game, UG) is as frequent in large-scale as in small-scale societies, while the occurrence of "altruistic" third-party punishment is mostly limited to the large-scale ones (Marlowe et al., 2008, 2011). Moreover, participants from the large societies tend to use more third-party punishment than spiteful second-party punishment. Marlowe et al. (2011) suggested that this distribution of the third- and second-party punishment points to the spiteful origins of human cooperation. An aversion to a personally-unfavorable unequal split, regardless of whether it is caused by fairness concerns or spiteful preferences, appears to be a human universal (see also Price, 2005).

If the long-term relationship between rank in a group and success as an individual is strong, then paying a small cost in order to acquire a higher rank by harming another individual may pay off in the long run. There are numerous examples in the animal world where the dominant

679 individual benefits disproportionately from being higher-ranked than the
680 second-highest individual in the hierarchy. Reproductive skew – that is,
681 the monopolizing of reproduction by alpha males and females – has been
682 observed in many species (e.g. Nelson-Flower et al., 2011; Setchell,
683 Charpentier, & Wickings, 2005; Sumner, Casiraghi, Foster, & Field,
684 2002). Rank may be particularly important in smaller groups in which it
685 is possible for one individual to control all potential competitors
686 (Kutsukake & Nunn, 2006).
687 In a situation where between-group competition is relatively low, the in-
688 group members become the main competitors for resources. In such
689 circumstances, one should expect indiscriminate punishment because
690 both altruistic and antisocial types of punishment increase the positive
691 payoff difference between the punisher and the punished. By Sylwester et
692 al.'s calculation (submitted), over 50% of participants from Muscat,
693 Athens, Samara and Riyadh in the Herrmann et al.'s (2008) study used
694 both antisocial and altruistic punishment over the course of ten rounds.
695 Both types of punishment were sometimes used within the same round[21].

[21] 11% of all punishment opportunities in Muscat and 9% in Riyadh showed mixed strategies. This is despite the fact that only half of the participants in the groups of four were able to punish this way on any given round, since by our definitions the highest contributors could not punish antisocially, nor the lowest altruistically.

In the data gathered by Herrmann et al. (2008), there is a negative relationship between GDP per capita (the measure of prosperity in a country) and the amount of antisocial punishment. GDP per capita is also highly correlated with the rule of law, used by the researchers as the main explanatory variable for antisocial punishment. Both the rule of law and antisocial punishment are constructs created to describe peoples' attitudes and behaviors. The correlation between the two is important but circular – it is difficult to infer causality. GDP per capita is interdependent with these characteristics but is also a measure describing the socio-ecology of a given place and defines its living conditions. A common finding in both biology and sociology is that as resources become scarcer, local competition between individuals increases (Briones, Montana, & Ezcurra, 1998; Grossman & Mendoza, 2003). In the context of enhanced local competition caused by waning resources, relative payoffs may matter more than absolute payoffs. In societies with high income-inequality and economic instability, the perceived risks caused by decreasing resource availability may maximize competitive predispositions and induce aggression towards in-group members.

Individual variation in antisocial punishment

Differences in punishment strategies also exist within groups from relatively homogenous populations. There are two possible explanations of individual variation in antisocial punishment in such

718 groups. Sanctioning cooperators could be a strategic behavior dependent
719 on the immediate circumstances, or it could constitute a relatively stable
720 part of an individual's personality. These two possibilities are not
721 exclusive – recent results indicate that both may be true.
722 Negative reciprocity – responding to harmful behavior with harm
723 (also known as revenge or retaliation) – is widespread in humans.
724 Evidence from UGs shows that, across the world, people would rather
725 give up their profits than allow their partner to take a disproportionately
726 large share (Henrich et al., 2005). Similarly, in PGGs, people are willing
727 to punish those who exploited them and, as a result, became better off
728 (Fehr & Gächter, 2002). In anther study, participants playing PGGs, who
729 were kept aware of the running-total earnings of fellow players,
730 contributed significantly less than those who knew both earnings and
731 contributions. These, in turn, contributed less than participants knowing
732 contributions only (Nikiforakis, 2010). Further, punishment increased
733 dramatically when both earnings and contributions were known in
734 comparison to the condition with known contributions only. Punishment
735 was not greater when only earnings were known, but it was also not less
736 (Nikiforakis, 2010).
737 Proximately, negative reciprocity results from the neurological
738 underpinnings of vengeance. Individuals who punish those who behave
739 unfairly derive satisfaction through the activation of reward circuits in
740 the brain (De Quervain et al., 2004). De Quervain et al. (2004)

741 implemented an experimental condition where the result of an unequal
742 split was due to chance, rather than to an intentional decision of their
743 partner. In this case, the majority of participants reported no desire to
744 punish and only three out of 14 participants sanctioned their partners by a
745 small amount. De Quervain et al.'s results may be indirectly applied to
746 antisocial punishment considering that costly punishment of cooperators
747 is, at least to some extent, motivated by revenge.

748 Herrmann et al. (2008) suggested that retaliation might be a
749 possible reason for antisocial punishment. In the majority of the
750 investigated subject pools, the amount of the received punishment is
751 positively related to the scale of antisocial punishment. However, the
752 design typically used in behavioral economic experiments on costly
753 punishment does not allow for pinpointing revenge. In a standard setting,
754 punishment is anonymous and participants are unaware of who punished
755 them (e.g. Egas & Riedl, 2008; Falk et al., 2005; Fehr & Gächter, 2002;
756 Herrmann et al., 2008). They also cannot see how much punishment
757 other individuals receive and, thusly, they cannot assess whether
758 sanctioning affects their contributions. Unless the punished individual is
759 the top contributor, they might expect that any punishment they receive is
760 "deserved" and may have come from a more cooperative person. In any
761 case, their revenge is blind: individuals can only try to guess who
762 punished them in the preceding rounds.

A few studies have investigated the consequences of revealing the identity of punishers and adding the possibility of targeted revenge to the design. In some conditions of the experiments of Denant-Boemont et al.'s (2007), Nikiforakis's (2008) and Cinyabuguma et al.'s (2006) after the first punishment stage participants were able to pay to reduce others' payoffs for a second time. Depending on the study, participants were provided with different information about the punishment decisions of others. In Denant-Boemont et al.'s (2007) study participants were either told all details about punishment decisions and the identities of the punishers (full information condition), only who punished them and by how much (revenge only condition) or information about how other players were punished (no revenge condition). In the "no revenge" condition, despite the extra punishment stage, participants' contributions remained stable and similar to those observed when no extra punishment opportunity was available. In contrast, when participants could target those who punished them in the past, in the "full" information and "revenge only" conditions another punishment stage resulted in a decrease in cooperation. While in the "no revenge condition", the amount contributed to the PGG above group average negatively correlated with received punishment, this was not the case when individuals could target those who punished them (full information and revenge only conditions), suggesting the occurrence of antisocial punishment.

Nikiforakis (2008) adopted a design similar to Denant-Boemont et al.'s "revenge only" condition, in that participants could only punish those who had just punished them. Antisocial punishment levels were similar in the condition where counter-punishment was possible and in the control standard condition with one round of punishment. However, when counter-punishment was enabled, both altruistic punishment and cooperation declined dramatically. In the counter-punishment stage, those who were punished antisocially were more likely to counter-punish than those who were punished because of their low contributions. In Cinyabuguma et al.'s (2006) experiment, participants learned how much punishment was assigned to individuals who contributed above, below or equal to the average of group contributions without knowing which specific individuals were punished and by how much. Here, the addition of another punishment stage did not result in participants lowering their contributions. Neither did it lead to a significant increase in contributions.

In all three studies, in conditions where participants could target those who punished them in the past, the extra punishment stage negatively affected contributions to the public good. In those cases, punishment following contributions was lower than in the control condition without the second punishment stage. Clearly, the fear of revenge, suppressed sanctioning behavior in the first punishment stage, which in turn reduced cooperation. However, in the second stage of

808 punishment, sanctioning occurred frequently and was directed to both
809 those who had previously punished altruistically and antisocially. In
810 conclusion, individuals who behave in an uncooperative way and are
811 subsequently punished, when given a chance, tend to retaliate. A
812 combination of anger and the lack of guilt were found to be the main
813 emotional causes of such negative reciprocity (Hopfensitz & Reuben,
814 2009).

815 Blind revenge is likely to be the motivation of some of the
816 punishment observed in Herrmann et al.'s study. However, instances of
817 punishing cooperators, though rare, occurred even after the first round of
818 the PGG (the first punishment opportunity), where negative reciprocity
819 can be excluded as a possible motive. In several studies, negative social
820 preferences have been examined in circumstances where no motive for
821 punishment existed. When participants of an experiment conducted in the
822 Netherlands could destroy the partner's money without the fear of
823 retaliation, they did so in 40% of decisions (Abbink & Sadrieh, 2009). In
824 another experiment with Ukrainian participants, the destruction rate more
825 than doubled, from around 11% to 25%, when the cause of destruction
826 was made obscure to the partner (Abbink & Herrmann, 2011). This
827 suggests that the way in which experiments are framed, combined with
828 enhanced anonymity, can have a dramatic impact on people's behavior.
829 The fact that cooperative behavior is often measured through experiments

that include an option to give, but not an option to take, may lead to biases in the interpretation of results.

Abbink and Sadrieh (2009) speculated that reducing another person's income even at one's own cost "gives pleasure". Such an interpretation is difficult to reconcile with the known "warm glow" effect caused by helping others (Andreoni, 1995) and the finding that contributing to the public good activates reward areas in the brain (Harbaugh, Mayr, & Burghart, 2007), though there is known to be individual variation in the level of such social rewards (Nettle, 2006). How can we then explain the high levels of "nastiness" observed in Abbink and Sadrieh's (2009) and Abbink and Herrmann's (2011) studies? It might be that rather than being pleasant, high levels of harming behaviour have been caused by the action bias, a preference to perform a given action rather than not do anything (Baron & Ritov, 2004; Patt & Zeckhauser, 2000). High rates of negative social behaviour might simply be an experimental fluke caused by the absence of any positive alternative. In a study where both costly rewards and costly punishment could be used, costly punishment almost disappeared while rewarding others remained at a stable high level over the course of rounds (Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009).[22]

[22] Interestingly, in Rand et al.'s (2009) experiment conducted in the U.S., unlike in other studies, punishment and reward decisions were not anonymous, so

A propensity for antisocial punishment may constitute part of a person's stable personality profile. In psychology, the Social Value Orientation (SVO) scale categorizes people with regard to how they value their personal payoff with reference to others' payoffs. A common finding is that the majority of participants (on average 46%) have, what SVO calls a "pro-social" orientation i.e. they choose that they and the other individuals receive an equal payoff (Au & Kwong, 2004). A smaller proportion of individuals (38%) choose the selfish option that maximizes their own absolute payoff. There is also an even smaller group (12%) that the SVO labels as "competitive". Competitive individuals favor a split that results in an increase in their own relative payoff, unlike selfish individuals who seek to maximize their absolute payoff.[23] While SVO may offer a proximate reason for why some individuals express antisocial punishment, it does not address the

---

participants could target those who affected their payoffs in the past. Despite this possibility of revenge (discussed in detail earlier in this section), punishment patterns resembled those observed in experiments with an anonymous design. Only a small amount of antisocial punishment occurred (see Rand et al.'s supplementary material). Herrmann et al. (2008) also reported very low levels of antisocial punishment in their only American city.

[23] In the SVO scale it is not possible to choose a distribution in which the other individual's payoff would be higher than own payoff.

evolutionary underpinnings of its distribution in a population. We will return to this topic in the next section.

Behavioral economics has also noted that social preferences are heterogeneous and that people can be classified into distinct types who behave in a relatively consistent and predictable manner (Fischbacher & Gächter, 2006; Gächter & Thöni, 2005; Kurzban & Houser, 2005). Their classification system is somewhat different than that adopted by social psychologists. The majority of individuals fall into the category called "conditional cooperators" or "reciprocators", that is, they are social learners who react to others' behavior. Due to their fine-tuning of behavior to free-riders' lack of cooperation, contributions in PGG decline over time (Fischbacher, Gächter, & Fehr, 2001). The two smaller groups are made up of cooperators who consistently act in a way that increases group welfare and free-riders who consistently pursue their own payoff maximizing interest.

The environment in which one develops may shape individual preferences for punishment behaviour. The choice of punishment type one imposes has been linked to the degree of discounting the future (Espin et al., 2012). In this study, conducted in Spain, present-oriented participants meted out more antisocial punishment and less altruistic punishment than their future-oriented counterparts. Discounting the future and focusing on present competition may be a successful strategy in unpredictable environments with scarce resources. In contrast,

887 enforcing cooperation with an expectation of future benefits is likely to
888 be a successful strategy in more stable and wealthy places (see Hill,
889 Jenkins & Farmer, 2008). Espin et al.'s (2012) results fit well with those
890 obtained by Herrmann et al. (2008), showing a negative correlation
891 between the expression of antisocial punishment and GDP per capita.
892 The notion that individuals' economic decisions in one game are
893 relatively stable and that they can be predictive of the decisions in
894 another game has been challenged by Herrmann and Orzen (2008).
895 Individuals classified as pro-social (altruists and conditional cooperators)
896 in a prisoner's dilemma problem, when presented with a contest game,
897 invested more aggressively than individuals classified as selfish.[24]
898 Moreover, individuals who played the contest game before, instead of
899 after, the prisoner's dilemma problem showed a decrease in cooperative
900 behavior. Herrmann and Orzen's results suggest that different game
901 contexts may shift individual social preferences; a "pro-social" type may
902 behave cooperatively in games framed as cooperative. When the game is
903 framed as competitive, their preference may reverse. The reduction in
904 cooperative behavior, after participation in a contest game, indicates that

[24] In the prisoner's dilemma problem, an individual who defects while their partner cooperates receives the highest payoff. The second highest payoff is when both partners cooperate. A lower payoff is obtained when both partners defect. The lowest payoff, the so called "sucker's payoff", is obtained by a person who cooperates while their partner defects.

905 the exposure to competitive situations and environments may
906 considerably affect the behavior of otherwise pro-social types[25].
907 When the possibility of punishment exists, social learners use it
908 and can achieve high levels of cooperation. Ones and Putterman (2007)
909 examined punishment behavior of individuals who were (unknowingly)
910 grouped according to their cooperative type[26]. Punishment patterns (no
911 punishment, altruistic punishment and antisocial punishment) remained
912 consistent across a number of rounds and were present even in the end
913 periods in which there were no incentives to punish, in terms of absolute
914 payoff. Antisocial punishers grouped together continued to punish
915 antisocially even in the final periods. Ones and Putterman's (2007)
916 results provide another piece of evidence indicating that the preferences
917 people hold cannot be narrowed down to absolute payoff maximization.
918 Importantly, they also suggest that antisocial punishment is not

---

[25] Note that this does not necessarily undermine the idea of individuals having stable strategies, rather it may mean the strategies are more complex than uniform pro- or antisocial behavior.

[26] The cooperative type was determined on the basis of five diagnostic rounds of PGG with punishment. After each round participants were reshuffeled between groups in a way to make the groups as diverse with respect to PGG contributions and punishment as possible. Next, participants were ranked according to their average contribution and punishment level in the five diagnostic rounds.

919 necessarily strategic and it may sometimes constitute a persistent
920 individual strategy.
921 Gächter and Thöni (2005) used a one-shot PGG in order to
922 determine participants' cooperative preferences. Participants who
923 contributed similar amounts of money in this diagnostic round were then
924 grouped together and showed the previous contributions of other group
925 members[27]. Hence, unlike in Ones and Putterman's (2007) design,
926 participants knew they would be interacting with like-minded people.
927 In the unsorted control condition the level of contributions in the
928 diagnostic one-shot PGG round differed considerably from the first
929 contribution round in the series of PGGs. This suggests that the prospect
930 of repeated interaction with people with similar strategies positively
931 affects behavior of all participants, including otherwise selfish
932 individuals. In the unsorted control condition, most punishment was
933 meted out by the lowest and the middle contributors but not by the
934 highest contributors. Participants from groups with the lowest
935 contributors meted out a considerable amount of antisocial punishment.
936 As in other studies, the type of cooperative preferences, determined

---

[27] Participants were ranked according to their contribution in the diagnostic round. The three top contributors formed one group, the next three highest the second group etc. For analysis, three classes of groups were created with the third of groups with the highest contributions, the third with the middle contributions and the third with the lowest contributions.

through the diagnostic round, remained consistent and affected punishment decisions. When participants knew that they were interacting with others of similar preferences, punishment by high and medium contributors almost disappeared (probably because both groups behaved in a very cooperative way) and the only punishing group were the lowest contributors. The information about whether antisocial punishment occurred in these sorted groups of low contributors is not provided.

As indicated above, motivations for antisocial punishment vary and do not necessarily involve revenge. At the most basic level, any instance of antisocial punishment is an expression of aggressive behavior (see Sylwester et al., submitted). Aggression may be used to undermine someone else's cooperative strategy or to defend one's own strategy. It may also result in gaining social status. In our view, costly antisocial punishment functions as a social signal to observers in the same way that altruistic acts do (Barclay, 2006; Hardy & van Vugt, 2006). By using antisocial punishment, individuals build a reputation for aggressiveness, which is likely to benefit them in some social contexts. It should be noted that while punishers may increase their payoff relative to the individual they punish, the cost of punishment means that they could also reduce their own payoff relative to that of non-punishing and unpunished individuals. By design, punishment is a costly game to play.

Antisocial punishment as an evolutionary strategy

When making evolutionary inferences based on behavioral economics experiments, it is important to take into account limitations and external validity of these experiments. Humans evolved in social groups where direct and indirect reciprocity played a role and it is likely that punishers could have been easily identified. Combining costly punishment with reputation can completely change the predicted evolutionary outcomes of different strategies (Santos, Rankin, & Wedekind, 2011). Contemporary large group size, anonymity and market integration may create circumstances resembling those present in behavioral economics experiments (e.g. online interactions). However, one needs to be cautious when extrapolating the results of such experiments to an evolutionary scale. In modern human societies, status is intrinsically related to cooperative reputation (Hardy & Van Vugt, 2006). When reputational information is public, as it was during the human evolutionary past, highly cooperative reputation facilitates the acquisition of desirable partners for profitable interactions (see e.g. Sylwester & Roberts, 2010). In the anonymous or pseudo-anonymous settings, used in behavioral economics experiments, at least some proportion of people may revert to the more basic way of establishing dominance – aggression.

Traulesen and colleagues (García & Traulsen, 2012; Hilbe & Traulsen, 2012) have recently used computer simulations to model the

evolutionary dynamics of reputation combined with sanctioning. They found that adding individual reputation into the simulation selected against all sanctioning, except that meted out to free riders (termed *altruistic* here). This may explain the difference between in-group and out-group behavior reported in the previous section – in-group individuals are, almost by definition, better known to group members than out-group ones. Therefore, the use of antisocial punishment may well vary between these conditions due to the availability of reputational information. It is worth noting, that the above models do not account for reputation gained from antisocial punishment. One can well imagine that an individual would adjust their behavior knowing that their partner tends to punish cooperators. Likewise, an uncooperative individual with a reputation for antisocial punishment might not receive much punishment from altruistic punishers because of a increased probability of retaliation.

It is possible that the high levels of antisocial punishment observed in some subject pools represent a sensible strategy under anonymous conditions. However, punishment that benefits the group can be viewed as a second order public good and can, therefore, improve reputation in non-anonymous settings. It has been shown that the presence of an audience enhances the use of third-party costly punishment against norm violators, even if that audience consists solely of the experimenter (Kurzban, DeScioli, & O'Brien, 2007). Investing in costly punishment that benefits the group is analogous to investing in

cooperation, and may positively affect reputation. Indeed, people who punish altruistically gain social benefits and higher earnings in paired interactions, thanks to their reputations as punishers (Barclay, 2006). Considering this strategic use of altruistic punishment, the high rates of antisocial punishment observed in several subject pools of Herrmann et al. (2008), may be manifested quite rarely in real life because of the reputational advantages of punishing free-riders. The small number of studies on reputation and punishment, and a lack of cross-cultural comparison of the effects of reputation, make prediction of the relationship between them difficult. In places with norms of low civic cooperation and weak rule of law, reputational benefits from altruistic punishment might not outweigh the benefits of the dominant status acquired by low contributions and antisocial punishment.

Evolutionary models show that even a small proportion of individuals with a particular strategy can have a dramatic effect on population dynamics. A simple example would be a small number of defectors who can invade a group of cooperators and make them disappear from the population (Maynard Smith, 1964, 1974). In a population where individuals use many different behavioral strategies, evolution may promote optimal mixes so that the local economic substrates are maximally exploited (MacLean, Fuentes-Hernandez, Greig, Hurst, & Gudelj, 2010; Nettle, 2006). Recently, agent-based modeling has been used to examine the consequences of adding

antisocial punishment to the repertoire of behaviors available to individuals in a society. In a simple model, the introduction of antisocial punishers led to the collapse of cooperation, and punishing antisocially became the dominant strategy (Rand, Armao, Joseph, Nakamaru, & Ohtsuki, 2010). In a population lacking a spatial structure costly punishment was evolutionarily stable. Punishers who could use both altruistic and antisocial punishment achieved the highest relative payoffs and eventually displaced non-punishers and punishers who specialized in one type of punishment. In a spatially structured population, defectors who did not punish and defectors who punished antisocially did best. In this case, antisocial punishment was a powerful strategy only rarely invaded by non-punishing defectors. In further models exploring the impact of group-structured populations due to Powers, Taylor and Bryson (2012) this result was showed to hold even in conditions of between-group competition. More generally, introducing antisocial punishment decreased the probability of the evolution of cooperation, though where group-level selection was sufficiently powerful (groups were small and persistent) cooperation could still evolve. Power et al.'s (2012) results indicate that antisocial punishment can only have evolved if it is inextricably associated with some other adaptive advantage, such as social dominance (see also Rand & Nowak, 2011). The evolutionary models summarised above lead to the conclusion that most of the mechanisms that have been proposed for the evolution of altruistic

1050 punishment can also promote antisocial punishment, if such strategies are
1051 not *a priori* excluded from the models.
1052 Costly punishment is usually modeled within the framework of
1053 the tragedy of the commons – despite initial cooperation, eventually all
1054 individuals become selfish payoff maximizers in repeated PGGs.
1055 However, many human interactions are likely to resemble not a tragedy
1056 of the commons but a tragedy of the commune (see Doebeli & Hauert,
1057 2005). Tragedy of the commune refers to a situation when the payoffs of
1058 cooperation and free-riding are based on the Snowdrift Game payoff
1059 matrix. In this game, mutual defection results in the worst possible payoff
1060 for both partners. An individual who defects in response to their partner's
1061 cooperation achieves the best possible payoff. In the tragedy of the
1062 commune cooperative types may co-exist with free-riders and
1063 cooperation can be maintained at a stable but low level (Doebeli &
1064 Hauert, 2005). Low but stable cooperation level was found by Herrmann
1065 et al. (2008) in subject pools with high antisocial punishment. If the
1066 payoff matrices of social dilemmas are more relaxed in real life than is
1067 assumed by a standard PGG, a mix of different cooperative types may be
1068 evolutionarily stable and therefore individuals might not be willing to use
1069 altruistic punishment to enforce cooperative norms.

2. Conclusions

In this article, we have examined the psychological and ecological causes of antisocial punishment at the individual, group, cultural and evolutionary levels. The experimental subjects typically used to investigate costly punishment in behavioral economics were originally heavily biased towards participants from democratic and relatively affluent places (Henrich et al., 2010). This has resulted in antisocial punishment being historically regarded as the "ugly step-sister" to altruistic punishment and treated as a rare phenomenon, not deserving of scientific attention. Thanks to the seminal study by Herrmann et al. (2008), we now know that, although rare in some contexts, in other contexts antisocial punishment constitutes a behavior as widely expressed as altruistic punishment. We have proposed that the contexts where antisocial punishment is pervasive may be the ones in which being locally competitive is likely to provide a considerable improvement in the socio-economic condition of the individual. In these contexts, cooperation remains stable, but it is at a lower level, relative to other regions. This is, possibly, because a small but stable proportion of individuals exhibit a preference for aggressive competition. Antisocial punishment is also more prevalent between individuals who do not consider each other "in-group". We have presented two explanations for this: both between-group competition, and selection against antisocial punishment in contexts where reputational cost is involved. Antisocial

1093 punishment therefore does not have to be viewed as an exceptionally
1094 complex or perplexing behavior. Rather, it can be easily described as
1095 aggression driven by competition (Sylwester, Mitchell &Bryson,
1096 submitted).
1097 As Darwin (1871) aptly put it, humans normally show extensive
1098 cooperation but in some circumstances their "lower, though at the
1099 moment, stronger impulses or desires" (p.104) may prevail. Recent
1100 reports concerning antisocial punishment have often emphasized the
1101 "dark side" of human nature, indicating such behavior is purely
1102 destructive. However, when viewed from an ecological perspective,
1103 punishing cooperators may be just one way to gain an advantage over
1104 others and may constitute a selfish behavior that positively affects
1105 individual survival and well-being. Costly punishment – whether
1106 altruistic or not – can be seen as a second-order public good because it
1107 may improve group cooperation and payoffs (Yamagishi, 1986). It can
1108 also be viewed as an effective weapon when used in individual
1109 competition.
1110 In addition to disputing that antisocial punishment is irrational,
1111 we have also disputed the hypothesis that costly punishment reliably acts
1112 as an independent mechanism for enhancing cooperation (Fehr &
1113 Gächter, 2002). Rather, when the opportunity to build reputation exists,
1114 punishment should be treated as a derivative of direct and indirect
1115 reciprocity. As Dreber et al. (2008) suggest, "costly punishment might

have evolved for reasons other than promoting cooperation, such as coercing individuals into submission and establishing dominance hierarchies" (p.350). Antisocial punishment is one example of such a mechanism.

We have shown that antisocial punishment, although initially costly to the punisher, may bring benefits in the long term (see Fig. 1). The circumstances favoring antisocial punishment are defined by the groups and cultures within which individuals are embedded. The evidence indicates that, at a micro-level, antisocial punishment often takes the form of negative reciprocity and may be a direct response to other individuals' behavior or that it is an expression of a competitive preference. Is *homo homini lupus*? Yes, if the ecological and cultural pressures make competitive behavior a successful strategy. However, with omnipresent reputation-based mechanisms of cooperation, which are not accounted for by behavioral economics experiments, such pressures are likely to be counteracted in ordinary real world interactions.

References

Abbink, K., Brandts, J., Herrmann, B., & Orzen, H. (2010). Intergroup conflict and intra-group punishment in an experimental contest game. *The American Economic Review*, *100*(1), 420–447.

Abbink, K., & Herrmann, B. (2011). The moral costs of nastiness. *Economic Inquiry*, *49*(2), 631–633.

Abbink, K., & Sadrieh, A. (2009). The pleasure of being nasty. *Economics Letters*, *105*(3), 306–308.

Anderson, C. M., & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, *54*(1), 1–24.

Andreoni, J. (1995). Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *The Quarterly Journal of Economics*, *110*(1), 1–21.

Au, W. T., & Kwong, J. Y. . (2004). Measurements and effects of social-value orientation in social dilemmas: A review. In *Contemporary Psychological Research on Social Dilemmas* (pp. 71–98). Cambridge University Press.

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

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74–85.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(7105), 912–915.

Binmore, K. (2007). *Does Game Theory Work? The Bargaining Challenge*. Cambridge, MA: MIT Press.

Briones, O., Montana, C., & Ezcurra, E. (1998). Competition intensity as a function of resource availability in a semiarid ecosystem. *Oecologia*, *116*(3), 365–372.

Carpenter, J. P. (2007). The demand for punishment. *Journal of Economic Behavior & Organization*, *62*(4), 522–542.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, *9*(3), 265–279.

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. New York: Appleton and Company.

De Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254 –1258.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic theory*, *33*(1), 145–167.

Doebeli, M., & Hauert, C. (2005). Models of cooperation based on the Prisoner's Dilemma and the Snowdrift game. *Ecology Letters*, *8*(7), 748–766.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*(7185), 348–351.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1637), 871.

Ellingsen, T., Herrmann, B., Nowak, M. A., Rand, D. G. and Tarnita, C. E. (2012). Civic capital in two cultures: The nature of cooperation in Romania and USA. Available at SSRN: http://ssrn.com/abstract=2179575

Espin, A. M., Brañas-Garza, P., Herrmann, B., & Gamella, J. F. (2012) Patient and impatient punishers of free-riders. *Proc Biol Sci. 279*, 4923-8.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *73*(6), 2017–2030.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*, 785–91.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Fehr, E., Hoff, K., & Kshetramade, M. (2008). Spite and development. *American Economic Review*, *98*, 494–499.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly journal of Economics*, *114*(3), 817–868.

Fischbacher, U., & Gächter, S. (2006). Heterogeneous social preferences and the dynamics of free riding in public goods. *IZA Discussion Paper 2011*.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 791.

Gächter, S., & Herrmann, B. (2011). The limits of self-governance in the presence of spite: Experimental evidence from urban and rural Russia. *European Economic Review*, *55*, 193–210.

Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), 2651.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*(5907), 1510.

Gächter, S., & Thöni, C. (2005). Social learning and voluntary cooperation among like-minded people. *Journal of the European Economic Association*, *3*(2-3), 303–314.

García, J., & Traulsen, A. (2012). Leaving the loners alone: evolution of cooperation in the presence of antisocial punishment. *Journal of theoretical biology*, *307*, 168–173.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*(2), 169–179.

Gintis, H., Bowles, S., Boyd, R. T., & Fehr, E. (2005). *Moral Sentiments and Material Interests: The Foundation of Cooperation in Economic Life*. Cambridge: The MIT Press.

Goette, L., Huffman, D., Meier, S., & Sutter, M. (2010). Group membership, competition, and altruistic versus antisocial punishment: Evidence from randomly assigned army groups. *IZA Discussion Paper 5189*.

Grossman, H. I., & Mendoza, J. (2003). Scarcity and appropriative competition. *European Journal of Political Economy*, *19*(4), 747–758.

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, *316*(5831), 1622.

Hardy, C. L., & Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin*, *32*(10), 1402–1413.

Henrich, J., Boyd, R. T., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (2004). *Foundations of Human Sociality: Economic Experiments*

and Ethnographic Evidence from Fifteen Small-Scale Societies. Oxford: Oxford University Press.

Henrich, J., Boyd, R. T., Bowles, S., Camerer, C., Fehr, E., Gintis, H., … Ensminger, J. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*(06), 795–815.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Henrich, N. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767.

Herrmann, B., & Orzen, H. (2008). The appearance of homo rivalis: Social preferences and the nature of rent seeking. *CeDEx discussion paper 2008-10, University of Nottingham*.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–403.

Hilbe, C., & Traulsen, A. (2012). Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. *Sci. Rep.*, *2*.

Hill, E. M., Jenkins, J., & Farmer, L. (2008). Family unpredictability, future discounting, and risk taking. *The Journal of Socio-Economics*, 37(4), 1381-1396.

Hoff, K., Kshetramade, M., & Fehr, E. (2009). Caste and Punishment. *Research Working papers*, *1*(1), 1–42.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Sage Publications, Inc.

Hopfensitz, A., & Reuben, E. (2009). The Importance of Emotions for the Effectiveness of Social Punishment. *The Economic Journal*, *119*(540), 1534–1559.

Houser, D., & Xiao, E. (2010). Inequality seeking punishment. *Economics Letters*, *109*, 20–23.

Inglehart, R., & Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, *65*(1), 19–51.

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), 2635.

Kümmerli, R., Burton-Chellew, M. N., Ross-Gillespie, A., & West, S. A. (2010). Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods

games. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(22), 10125–10130.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84.

Kurzban, R., & Houser, D. (2005). Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(5), 1803–1807.

Kutsukake, N., & Nunn, C. L. (2006). Comparative tests of reproductive skew in male primates: the roles of demographic factors and incomplete control. *Behavioral Ecology and Sociobiology*, *60*(5), 695–706.

Leach, C. W., Spears, R., Branscombe, N. R., & Doosje, B. (2003). Malicious pleasure: Schadenfreude at the suffering of another group. *Journal of Personality and Social Psychology*, *84*(5), 932–943.

MacLean, R. C., Fuentes-Hernandez, A., Greig, D., Hurst, L. D., & Gudelj, I. (2010). A Mixture of "Cheats" and "Co-Operators" Can Enable Maximal Group Benefit. *PLoS Biology*, *8*(9).

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., … Tracer, D. (2008). More "altruistic"

punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1634), 587 –592.

Marlowe, F. W., Berbesque, J. C., Barrett, C., Bolyanatz, A., Gurven, M., & Tracer, D. (2011). The "spiteful" origins of human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *278*, 2159–2164.

Masclet, D., Noussair, C., Tucker, S., & Villeval, M. C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review*, *93*(1), 366–380.

Maynard Smith, J. (1964). Group selection and kin selection. *Nature*, *201*(4924), 1145–1147.

Maynard Smith, J. (1974). The Theory of Games and the Evolution of Animal Conflicts. *Journal of Theoretical Biology*, *47*, 209–21.

McLeish, K. N., & Oxoby, R. J. (2007). Identity, cooperation, and punishment. *IZA Discussion Paper 2572*.

Nelson-Flower, M. J., Hockey, P. A. ., O'Ryan, C., Raihani, N. J., Du Plessis, M. A., & Ridley, A. R. (2011). Monogamous dominant pairs monopolize reproduction in the cooperatively breeding pied babbler. *Behavioral Ecology*, *22*(3), 559.

Nettle, D. (2006). The Evolution of Personality Variation in Humans and Other Animals. *American Psychologist*, *61*(6), 622–631.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, *92*(1-2), 91–112.

Nikiforakis, N. (2010). Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior*, *68*(2), 689–702.

Ones, U., & Putterman, L. (2007). The ecology of collective action: A public goods and sanctions experiment with controlled group formation. *Journal of Economic Behavior & Organization*, *62*(4), 495–521.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *The American Political Science Review*, *86*(2), 404–417.

Page, T., Putterman, L., & Garcia, B. (2008). Getting punishment right: Do costly monitoring or redistributive punishment help? *Brown University discussion paper 2008-1*.

Patt, A., & Zeckhauser, R. (2000). Action bias and environmental decisions. *Journal of Risk and Uncertainty*, *21*(1), 45–72.

Powers, S. T., Taylor, D. J., & Bryson, J. J. (2012). Punishment can promote defection in group-structured populations. *Journal of Theoretical Biology*, *311*(0), 107–116.

Price, M. E. (2005). Punitive sentiment among the Shuar and in industrialized societies: Cross-cultural similarities. *Evolution and Human Behavior*, *26*(3), 279–287.

Queller, D., C. (1994). Genetic relatedness in viscous populations. *Evolutionary Ecology*, *8*(1), 70–73.

Rand, D. G., Armao, I. V., Joseph, J., Nakamaru, M., & Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology*.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science*, *325*(5945), 1272–1275.

Rand, D. G., & Nowak, M. A. (2011). The evolution of antisocial punishment in optional public goods games. *Nat Commun*, *2*, 434.

Sääksvuori, L., Mappes, T., & Puurtinen, M. (2011). Costly punishment prevails in intergroup conflict. *Proceedings of the Royal Society B: Biological Sciences*.

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

Santos, M. dos, Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1704), 371 –377.

1378 Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary
1379 Theory and the Ultimate–Proximate Distinction in the Human
1380 Behavioral Sciences. *Perspectives on Psychological Science*,
1381 *6*(1), 38–47. doi:10.1177/1745691610393528
1382 Setchell, J. M., Charpentier, M., & Wickings, E. J. (2005). Mate guarding
1383 and paternity in mandrills: factors influencing alpha male
1384 monopoly. *Animal Behaviour*, *70*(5), 1105–1120.
1385 Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are
1386 worse than bitter enemies: "Altruistic" punishment of in-group
1387 members. *Evolution and Human Behavior*, *25*(6), 379–393.
1388 Sumner, S., Casiraghi, M., Foster, W., & Field, J. (2002). High
1389 reproductive skew in tropical hover wasps. *Proceedings of the*
1390 *Royal Society of London. Series B: Biological Sciences*,
1391 *269*(1487), 179.
1392 Sylwester, K., Mitchell, J., & Bryson, J. J. (in preparation). Punishment
1393 as aggression: Uses and consequences of costly punishment
1394 across populations.
1395 Sylwester, K., & Roberts, G. (2010). Cooperators benefit through
1396 reputation-based partner choice in economic games. *Biology*
1397 *Letters*, *6*(5), 659.
1398 Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific*
1399 *American*, *223*(5), 96–102.

Thierry, B. (2005). Integrating proximate and ultimate causation: Just one more go! *Current Science*, *89*, 1180–1183.

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), 415–432.

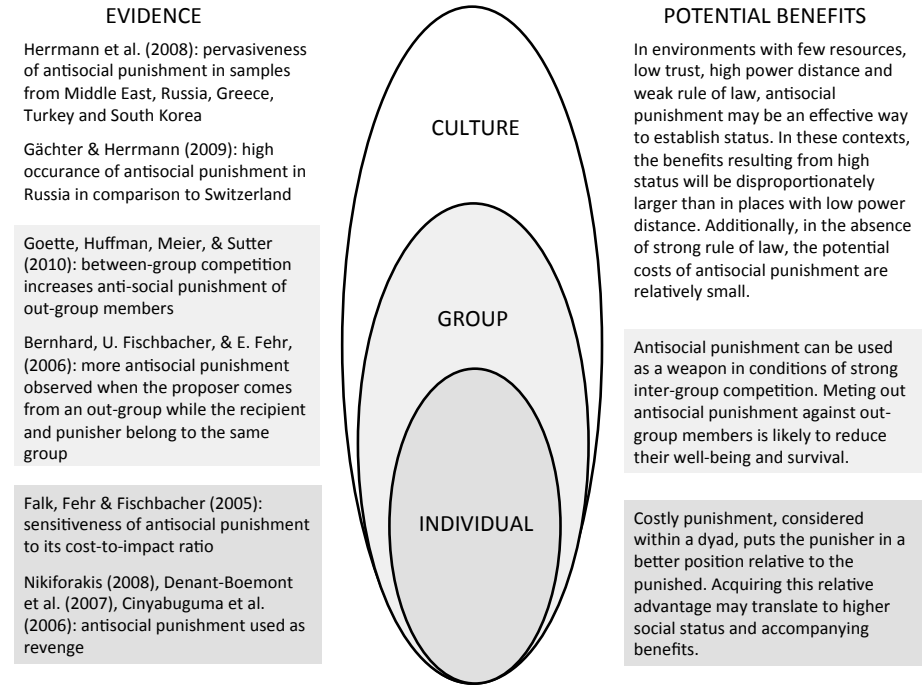
Wilson, D. S. (2004). What is wrong with absolute individual fitness? *Trends in Ecology & Evolution*, *19*(5), 245–248.

Wu, J.-J., Zhang, B.-Y., Zhou, Z.-X., He, Q.-Q., Zheng, X.-D., Cressman, R., & Tao, Y. (2009). Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences*, *106*(41), 17448 –17451.

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

1414 Figure 1 Antisocial punishment at individual, group and cultural level

1415 with its possible benefits



1416

Table 1

Different Consequences of Costly Punishment in the PGG

	Stage	P1	P2	P3	P4
Round 1	PGG contribution	2	4	10	20
	Punishment decision	P2	-	P1	P2
Round 2	PGG contribution	4	6	10	18

P4's behavior is a classical example of altruistic punishment. A cooperative individual P4 suffers a cost to punish P2 who contributed less than the group average. As a result of this punishment, P2's contribution increases in the next PGG round. Consider the behavior of P1 who punished P2. As a result of the punishment, P2 increased their contributions. Therefore, P1's punishment can be called functionally altruistic. At the same time, this punishment would be defined as antisocial (*sensu* Herrmann et al., 2008) because P1's original contribution, which is lower than P2's contribution, is treated as a reference level.