



*Citation for published version:*

Sylwester, K, Herrmann, B & Bryson, J 2013, 'Homo homini lupus? Explaining antisocial punishment', *Journal of Neuroscience, Psychology, and Economics*, vol. 6, no. 3, pp. 167-188. <https://doi.org/10.1037/npe0000009>

*DOI:*

[10.1037/npe0000009](https://doi.org/10.1037/npe0000009)

*Publication date:*

2013

*Document Version*

Peer reviewed version

[Link to publication](#)

© APA. This article may not exactly replicate the final version published in the APA journal. It is not the copy of record.

## University of Bath

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Homo homini lupus?* Explaining antisocial punishment

Karolina Sylwester<sup>1</sup>

Benedikt Herrmann<sup>2</sup>

Joanna J. Bryson<sup>1</sup>

<sup>1</sup> University of Bath

Department of Computer Science

BA2 7AY Bath

United Kingdom

<sup>2</sup> Behavioural Economics Team

Joint Research Centre

European Commission

## EXPLAINING ANTISOCIAL PUNISHMENT

## 1 ABSTRACT

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

22 ***Homo homini lupus? Explaining antisocial punishment***

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

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

---

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

<sup>2</sup> (Dreber, Rand, Fudenberg, & Nowak, 2008; Wu et al., 2009)

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

## EXPLAINING ANTISOCIAL PUNISHMENT

61 its functional causes.<sup>3</sup> In addition to proposing and justifying this  
62 theoretical framework, we also emphasize some unresolved questions  
63 about costly punishment, and offer testable predictions.

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

---

<sup>3</sup> 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.

## EXPLAINING ANTISOCIAL PUNISHMENT

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

**94 Costly punishment terminology**

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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



## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

---

<sup>4</sup> 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).

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

## EXPLAINING ANTISOCIAL PUNISHMENT

161 punishment behavior should ever be considered “altruistic”, in the folk-  
162 psychological sense.

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

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

---

<sup>5</sup> 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.”

<sup>6</sup> 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.

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

### 211 **1. The price of punishment**

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

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

## EXPLAINING ANTISOCIAL PUNISHMENT

222 as the punishment price increases (e.g. Anderson & Putterman, 2006).  
223 Despite this, some costly punishment (mostly directed at uncooperative  
224 individuals) is observed even when the cost to the punisher is larger than  
225 the cost to the punished individual. Antisocial punishment does occur,  
226 though rarely, even with a high relative cost of punishment (Anderson &  
227 Putterman, 2006).<sup>7</sup>

228           There is variation in the results reported concerning sensitivity to  
229 the relative cost of punishment. Using data from U.S. participants,  
230 Carpenter (2007) analyzed the behavior of free-riders who punished  
231 cooperators, cooperators who punished free-riders and free-riders who  
232 punished other free-riders.<sup>8</sup> Out of the three groups, free-riders punishing  
233 other free-riders were most sensitive to the price of punishment. Free-  
234 riders who punished cooperators did not condition their punishment  
235 decisions on price. Carpenter's results contrast with those obtained by  
236 Falk, Fehr & Fischbacher who used a sample of Swiss participants  
237 (2005). These researchers found that when the cost of punishment is the  
238 same to the punisher as to the punished, antisocial punishment

---

<sup>7</sup> 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.

<sup>8</sup> 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.

## EXPLAINING ANTISOCIAL PUNISHMENT

239 disappears.<sup>9</sup> When punishment resulted in lowering the payoff of the  
240 punished person to a greater extent than reducing the cost to the punisher,  
241 sanctioning of cooperators by defectors and defectors by other defectors  
242 occurred frequently.

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

---

<sup>9</sup> 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.

<sup>10</sup> The price-to-impact ratios used by Egas and Riedl were: 1/3, 3/1, 1/1 and 3/3

<sup>11</sup> The researchers call this counter-intuitive punishment.

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

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



## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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

**313 Cross-cultural variation in punishment**

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

## EXPLAINING ANTISOCIAL PUNISHMENT

324 from such a debt. In consequence, cooperators were punished  
325 antisocially, but, interestingly, in this situation the cost to the punisher  
326 was even higher than the cost to the punished.<sup>12</sup>

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

---

<sup>12</sup> 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).

<sup>13</sup> Scandinavia, the UK, Germany & Switzerland. Southwestern Europe, e.g. France, Spain & Italy were not tested.

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

---

<sup>14</sup> 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).

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

---

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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 Herrmann 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

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

---

<sup>16</sup> The reference level was the group average.



## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

## EXPLAINING ANTISOCIAL PUNISHMENT

467 antisocial punishment.<sup>17</sup> 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

---

<sup>17</sup> Though voluntary organisations in the former Soviet Union might have a different character from voluntary organisations in established market economies.

## EXPLAINING ANTISOCIAL PUNISHMENT

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

**495 Antisocial punishment at the group level**

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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

511 ***Inter-group competition***

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

## EXPLAINING ANTISOCIAL PUNISHMENT

530 group members and fall victim of antisocial punishment.<sup>18</sup>

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

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

547           Indeed, experiments conducted in Papua New Guinea with two

---

<sup>18</sup> 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.

## EXPLAINING ANTISOCIAL PUNISHMENT

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).<sup>19</sup> 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),

---

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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, &

## EXPLAINING ANTISOCIAL PUNISHMENT

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

604         The levels of antisocial punishment observed in conventional  
605 PGG experiments appear to be low when contrasted with the levels  
606 towards the out-group members induced by conflict. A possible  
607 interpretation of this would be that punishment in ordinary PGG is only a  
608 side-effect of mechanisms evolved for conflict situations. The act of  
609 costly punishment, when taken out of the PGG context, can be perceived  
610 as mere aggression. Engaging in aggressive interactions with out-group  
611 members in the presence of conflict may be advantageous, in that it may  
612 help preserve a group's resources such as territory. Herrmann et al.  
613 (2008) found a negative correlation between antisocial punishment and



## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

628 ***Intra-group competition***

629 In ecological contexts where intra-group competition is fierce,  
630 individuals will use aggression towards members of their own group.  
631 Costly punishment typically decreases average payoffs (Dreber et al.,  
632 2008; Wu et al., 2009), however, it might be useful for displaying  
633 aggression and gaining relative advantage over the punished individual.  
634 Previous research has shown that people do care about their relative  
635 payoff within a group. For example, Saijo and Nakamura (1995) made  
636 participants face a non-dilemma in which the payoff maximizing choice

## EXPLAINING ANTISOCIAL PUNISHMENT

637 was to contribute the whole allocation to the public pool<sup>20</sup>. Although the  
638 contributions to PGG in the non-dilemma condition were higher than in  
639 the standard dilemma, a considerable number of participants still  
640 refrained from contributing and failed to maximize their payoff. The  
641 average amounts saved in the non-dilemma situation were higher than the  
642 average investments to the public pool in the traditional dilemma version  
643 of the PGG. This indicates that in the no-dilemma situation more  
644 participants chose the non-optimal (non-payoff-maximizing) outcome  
645 than in the traditional dilemma, indicating that the non-dilemma may  
646 have been taken as a spiteful dilemma.

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

---

<sup>20</sup> 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.

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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<sup>21</sup>.

---

<sup>21</sup> 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.

## EXPLAINING ANTISOCIAL PUNISHMENT

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

**714 Individual variation in antisocial punishment**

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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 another 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)

## EXPLAINING ANTISOCIAL PUNISHMENT

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.

## EXPLAINING ANTISOCIAL PUNISHMENT

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



## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

---

<sup>22</sup> 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

## EXPLAINING ANTISOCIAL PUNISHMENT

850           A propensity for antisocial punishment may constitute part of a  
851 person's stable personality profile. In psychology, the Social Value  
852 Orientation (SVO) scale categorizes people with regard to how they  
853 value their personal payoff with reference to others' payoffs. A common  
854 finding is that the majority of participants (on average 46%) have, what  
855 SVO calls a "pro-social" orientation i.e. they choose that they and the  
856 other individuals receive an equal payoff (Au & Kwong, 2004). A  
857 smaller proportion of individuals (38%) choose the selfish option that  
858 maximizes their own absolute payoff. There is also an even smaller  
859 group (12%) that the SVO labels as "competitive". Competitive  
860 individuals favor a split that results in an increase in their own relative  
861 payoff, unlike selfish individuals who seek to maximize their absolute  
862 payoff.<sup>23</sup> While SVO may offer a proximate reason for why some  
863 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.

<sup>23</sup> 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.

## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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.<sup>24</sup>  
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

---

<sup>24</sup> 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.

## EXPLAINING ANTISOCIAL PUNISHMENT

905 the exposure to competitive situations and environments may  
906 considerably affect the behavior of otherwise pro-social types<sup>25</sup>.

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<sup>26</sup>. 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

---

<sup>25</sup> 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.

<sup>26</sup> The cooperative type was determined on the basis of five diagnostic rounds of PGG with punishment. After each round participants were reshuffled 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.

## EXPLAINING ANTISOCIAL PUNISHMENT

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<sup>27</sup>. 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

---

<sup>27</sup> 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.



## EXPLAINING ANTISOCIAL PUNISHMENT

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

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

**958 Antisocial punishment as an evolutionary strategy**

959           When making evolutionary inferences based on behavioral  
960 economics experiments, it is important to take into account limitations  
961 and external validity of these experiments. Humans evolved in social  
962 groups where direct and indirect reciprocity played a role and it is likely  
963 that punishers could have been easily identified. Combining costly  
964 punishment with reputation can completely change the predicted  
965 evolutionary outcomes of different strategies (Santos, Rankin, &  
966 Wedekind, 2011). Contemporary large group size, anonymity and market  
967 integration may create circumstances resembling those present in  
968 behavioral economics experiments (e.g. online interactions). However,  
969 one needs to be cautious when extrapolating the results of such  
970 experiments to an evolutionary scale. In modern human societies, status  
971 is intrinsically related to cooperative reputation (Hardy & Van Vugt,  
972 2006). When reputational information is public, as it was during the  
973 human evolutionary past, highly cooperative reputation facilitates the  
974 acquisition of desirable partners for profitable interactions (see e.g.  
975 Sylwester & Roberts, 2010). In the anonymous or pseudo-anonymous  
976 settings, used in behavioral economics experiments, at least some  
977 proportion of people may revert to the more basic way of establishing  
978 dominance – aggression.

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

## EXPLAINING ANTISOCIAL PUNISHMENT

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

995         It is possible that the high levels of antisocial punishment  
996 observed in some subject pools represent a sensible strategy under  
997 anonymous conditions. However, punishment that benefits the group can  
998 be viewed as a second order public good and can, therefore, improve  
999 reputation in non-anonymous settings. It has been shown that the  
1000 presence of an audience enhances the use of third-party costly  
1001 punishment against norm violators, even if that audience consists solely  
1002 of the experimenter (Kurzban, DeScioli, & O'Brien, 2007). Investing in  
1003 costly punishment that benefits the group is analogous to investing in

## EXPLAINING ANTISOCIAL PUNISHMENT

1004 cooperation, and may positively affect reputation. Indeed, people who  
1005 punish altruistically gain social benefits and higher earnings in paired  
1006 interactions, thanks to their reputations as punishers (Barclay, 2006).  
1007 Considering this strategic use of altruistic punishment, the high rates of  
1008 antisocial punishment observed in several subject pools of Herrmann et  
1009 al. (2008), may be manifested quite rarely in real life because of the  
1010 reputational advantages of punishing free-riders. The small number of  
1011 studies on reputation and punishment, and a lack of cross-cultural  
1012 comparison of the effects of reputation, make prediction of the  
1013 relationship between them difficult. In places with norms of low civic  
1014 cooperation and weak rule of law, reputational benefits from altruistic  
1015 punishment might not outweigh the benefits of the dominant status  
1016 acquired by low contributions and antisocial punishment.

1017         Evolutionary models show that even a small proportion of  
1018 individuals with a particular strategy can have a dramatic effect on  
1019 population dynamics. A simple example would be a small number of  
1020 defectors who can invade a group of cooperators and make them  
1021 disappear from the population (Maynard Smith, 1964, 1974). In a  
1022 population where individuals use many different behavioral strategies,  
1023 evolution may promote optimal mixes so that the local economic  
1024 substrates are maximally exploited (MacLean, Fuentes-Hernandez,  
1025 Greig, Hurst, & Gudelj, 2010; Nettle, 2006). Recently, agent-based  
1026 modeling has been used to examine the consequences of adding

## EXPLAINING ANTISOCIAL PUNISHMENT

1027 antisocial punishment to the repertoire of behaviors available to  
1028 individuals in a society. In a simple model, the introduction of antisocial  
1029 punishers led to the collapse of cooperation, and punishing antisocially  
1030 became the dominant strategy (Rand, Armao, Joseph, Nakamaru, &  
1031 Ohtsuki, 2010). In a population lacking a spatial structure costly  
1032 punishment was evolutionarily stable. Punishers who could use both  
1033 altruistic and antisocial punishment achieved the highest relative payoffs  
1034 and eventually displaced non-punishers and punishers who specialized in  
1035 one type of punishment. In a spatially structured population, defectors  
1036 who did not punish and defectors who punished antisocially did best. In  
1037 this case, antisocial punishment was a powerful strategy only rarely  
1038 invaded by non-punishing defectors. In further models exploring the  
1039 impact of group-structured populations due to Powers, Taylor and  
1040 Bryson (2012) this result was showed to hold even in conditions of  
1041 between-group competition. More generally, introducing antisocial  
1042 punishment decreased the probability of the evolution of cooperation,  
1043 though where group-level selection was sufficiently powerful (groups  
1044 were small and persistent) cooperation could still evolve. Power et al.'s  
1045 (2012) results indicate that antisocial punishment can only have evolved  
1046 if it is inextricably associated with some other adaptive advantage, such  
1047 as social dominance (see also Rand & Nowak, 2011). The evolutionary  
1048 models summarised above lead to the conclusion that most of the  
1049 mechanisms that have been proposed for the evolution of altruistic

## EXPLAINING ANTISOCIAL PUNISHMENT

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.

**1070 2. Conclusions**

1071           In this article, we have examined the psychological and  
1072 ecological causes of antisocial punishment at the individual, group,  
1073 cultural and evolutionary levels. The experimental subjects typically used  
1074 to investigate costly punishment in behavioral economics were originally  
1075 heavily biased towards participants from democratic and relatively  
1076 affluent places (Henrich et al., 2010). This has resulted in antisocial  
1077 punishment being historically regarded as the “ugly step-sister” to  
1078 altruistic punishment and treated as a rare phenomenon, not deserving of  
1079 scientific attention. Thanks to the seminal study by Herrmann et al.  
1080 (2008), we now know that, although rare in some contexts, in other  
1081 contexts antisocial punishment constitutes a behavior as widely  
1082 expressed as altruistic punishment. We have proposed that the contexts  
1083 where antisocial punishment is pervasive may be the ones in which being  
1084 locally competitive is likely to provide a considerable improvement in  
1085 the socio-economic condition of the individual. In these contexts,  
1086 cooperation remains stable, but it is at a lower level, relative to other  
1087 regions. This is, possibly, because a small but stable proportion of  
1088 individuals exhibit a preference for aggressive competition. Antisocial  
1089 punishment is also more prevalent between individuals who do not  
1090 consider each other “in-group”. We have presented two explanations for  
1091 this: both between-group competition, and selection against antisocial  
1092 punishment in contexts where reputational cost is involved. Antisocial

## EXPLAINING ANTISOCIAL PUNISHMENT

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



## EXPLAINING ANTISOCIAL PUNISHMENT

1116 have evolved for reasons other than promoting cooperation, such as  
1117 coercing individuals into submission and establishing dominance  
1118 hierarchies” (p.350). Antisocial punishment is one example of such a  
1119 mechanism.

1120         We have shown that antisocial punishment, although initially  
1121 costly to the punisher, may bring benefits in the long term (see Fig. 1).  
1122 The circumstances favoring antisocial punishment are defined by the  
1123 groups and cultures within which individuals are embedded. The  
1124 evidence indicates that, at a micro-level, antisocial punishment often  
1125 takes the form of negative reciprocity and may be a direct response to  
1126 other individuals’ behavior or that it is an expression of a competitive  
1127 preference. Is *homo homini lupus*? Yes, if the ecological and cultural  
1128 pressures make competitive behavior a successful strategy. However,  
1129 with omnipresent reputation-based mechanisms of cooperation, which  
1130 are not accounted for by behavioral economics experiments, such  
1131 pressures are likely to be counteracted in ordinary real world interactions.

1132

1133

1134 **References**

- 1135 Abbink, K., Brandts, J., Herrmann, B., & Orzen, H. (2010). Intergroup  
1136 conflict and intra-group punishment in an experimental contest  
1137 game. *The American Economic Review*, *100*(1), 420–447.
- 1138 Abbink, K., & Herrmann, B. (2011). The moral costs of nastiness.  
1139 *Economic Inquiry*, *49*(2), 631–633.
- 1140 Abbink, K., & Sadrieh, A. (2009). The pleasure of being nasty.  
1141 *Economics Letters*, *105*(3), 306–308.
- 1142 Anderson, C. M., & Putterman, L. (2006). Do non-strategic sanctions  
1143 obey the law of demand? The demand for punishment in the  
1144 voluntary contribution mechanism. *Games and Economic  
1145 Behavior*, *54*(1), 1–24.
- 1146 Andreoni, J. (1995). Warm-glow versus cold-prickle: the effects of  
1147 positive and negative framing on cooperation in experiments. *The  
1148 Quarterly Journal of Economics*, *110*(1), 1–21.
- 1149 Au, W. T., & Kwong, J. Y. . (2004). Measurements and effects of social-  
1150 value orientation in social dilemmas: A review. In *Contemporary  
1151 Psychological Research on Social Dilemmas* (pp. 71–98).  
1152 Cambridge University Press.
- 1153 Barclay, P. (2006). Reputational benefits for altruistic punishment.  
1154 *Evolution and Human Behavior*, *27*, 325–344.

## EXPLAINING ANTISOCIAL PUNISHMENT

- 1155 Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and  
1156 normality. *Organizational Behavior and Human Decision*  
1157 *Processes*, 94(2), 74–85.
- 1158 Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in  
1159 humans. *Nature*, 442(7105), 912–915.
- 1160 Binmore, K. (2007). *Does Game Theory Work? The Bargaining*  
1161 *Challenge*. Cambridge, MA: MIT Press.
- 1162 Briones, O., Montana, C., & Ezcurra, E. (1998). Competition intensity as  
1163 a function of resource availability in a semiarid ecosystem.  
1164 *Oecologia*, 116(3), 365–372.
- 1165 Carpenter, J. P. (2007). The demand for punishment. *Journal of*  
1166 *Economic Behavior & Organization*, 62(4), 522–542.
- 1167 Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order  
1168 punishment deter perverse punishment? *Experimental Economics*,  
1169 9(3), 265–279.
- 1170 Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*.  
1171 New York: Appleton and Company.
- 1172 De Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M.,  
1173 Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of  
1174 altruistic punishment. *Science*, 305(5688), 1254–1258.
- 1175 Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment,  
1176 counterpunishment and sanction enforcement in a social dilemma  
1177 experiment. *Economic theory*, 33(1), 145–167.

## EXPLAINING ANTISOCIAL PUNISHMENT

- 1178 Doebeli, M., & Hauert, C. (2005). Models of cooperation based on the  
1179 Prisoner's Dilemma and the Snowdrift game. *Ecology Letters*,  
1180 8(7), 748–766.
- 1181 Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008).  
1182 Winners don't punish. *Nature*, 452(7185), 348–351.
- 1183 Egas, M., & Riedl, A. (2008). The economics of altruistic punishment  
1184 and the maintenance of cooperation. *Proceedings of the Royal  
1185 Society B: Biological Sciences*, 275(1637), 871.
- 1186 Ellingsen, T., Herrmann, B., Nowak, M. A., Rand, D. G. and Tarnita, C.  
1187 E. (2012). Civic capital in two cultures: The nature of cooperation  
1188 in Romania and USA. Available at SSRN:  
1189 <http://ssrn.com/abstract=2179575>
- 1190 Espin, A. M., Brañas-Garza, P., Herrmann, B., & Gamella, J. F. (2012)  
1191 Patient and impatient punishers of free-riders. *Proc Biol Sci.* 279,  
1192 4923-8.
- 1193 Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind  
1194 informal sanctions. *Econometrica*, 73(6), 2017–2030.
- 1195 Fehr, E., & Fischbacher, U. (2003). The nature of human altruism.  
1196 *Nature*, 425, 785–91.
- 1197 Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*,  
1198 415(6868), 137–140.
- 1199 Fehr, E., Hoff, K., & Kshetramade, M. (2008). Spite and development.  
1200 *American Economic Review*, 98, 494–499.

## EXPLAINING ANTISOCIAL PUNISHMENT

- 1201 Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and  
1202 cooperation. *Quarterly journal of Economics*, 114(3), 817–868.
- 1203 Fischbacher, U., & Gächter, S. (2006). Heterogeneous social preferences  
1204 and the dynamics of free riding in public goods. *IZA Discussion*  
1205 *Paper 2011*.
- 1206 Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally  
1207 cooperative? Evidence from a public goods experiment.  
1208 *Economics Letters*, 71(3), 397–404.
- 1209 Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human  
1210 cooperation: previous insights and a new cross-cultural  
1211 experiment. *Philosophical Transactions of the Royal Society B:*  
1212 *Biological Sciences*, 364(1518), 791.
- 1213 Gächter, S., & Herrmann, B. (2011). The limits of self-governance in the  
1214 presence of spite: Experimental evidence from urban and rural  
1215 Russia. *European Economic Review*, 55, 193–210.
- 1216 Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation.  
1217 *Philosophical Transactions of the Royal Society B: Biological*  
1218 *Sciences*, 365(1553), 2651.
- 1219 Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of  
1220 punishment. *Science*, 322(5907), 1510.
- 1221 Gächter, S., & Thöni, C. (2005). Social learning and voluntary  
1222 cooperation among like-minded people. *Journal of the European*  
1223 *Economic Association*, 3(2-3), 303–314.

## EXPLAINING ANTISOCIAL PUNISHMENT

- 1224 García, J., & Traulsen, A. (2012). Leaving the loners alone: evolution of  
1225 cooperation in the presence of antisocial punishment. *Journal of*  
1226 *theoretical biology*, 307, 168–173.
- 1227 Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of*  
1228 *Theoretical Biology*, 206(2), 169–179.
- 1229 Gintis, H., Bowles, S., Boyd, R. T., & Fehr, E. (2005). *Moral Sentiments*  
1230 *and Material Interests: The Foundation of Cooperation in*  
1231 *Economic Life*. Cambridge: The MIT Press.
- 1232 Goette, L., Huffman, D., Meier, S., & Sutter, M. (2010). Group  
1233 membership, competition, and altruistic versus antisocial  
1234 punishment: Evidence from randomly assigned army groups. *IZA*  
1235 *Discussion Paper 5189*.
- 1236 Grossman, H. I., & Mendoza, J. (2003). Scarcity and appropriative  
1237 competition. *European Journal of Political Economy*, 19(4), 747–  
1238 758.
- 1239 Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses  
1240 to taxation and voluntary giving reveal motives for charitable  
1241 donations. *Science*, 316(5831), 1622.
- 1242 Hardy, C. L., & Van Vugt, M. (2006). Nice guys finish first: The  
1243 competitive altruism hypothesis. *Personality and Social*  
1244 *Psychology Bulletin*, 32(10), 1402–1413.
- 1245 Henrich, J., Boyd, R. T., Bowles, S., Camerer, C., Fehr, E., & Gintis, H.  
1246 (2004). *Foundations of Human Sociality: Economic Experiments*

## EXPLAINING ANTISOCIAL PUNISHMENT

- 1247            *and Ethnographic Evidence from Fifteen Small-Scale Societies.*  
1248            Oxford: Oxford University Press.
- 1249    Henrich, J., Boyd, R. T., Bowles, S., Camerer, C., Fehr, E., Gintis, H., ...  
1250            Ensminger, J. (2005). "Economic man" in cross-cultural  
1251            perspective: Behavioral experiments in 15 small-scale societies.  
1252            *Behavioral and Brain Sciences*, 28(06), 795–815.
- 1253    Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people  
1254            in the world? *Behavioral and Brain Sciences*, 33(2-3), 61–83.
- 1255    Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C.,  
1256            Bolyanatz, A., ... Henrich, N. (2006). Costly punishment across  
1257            human societies. *Science*, 312(5781), 1767.
- 1258    Herrmann, B., & Orzen, H. (2008). The appearance of homo rivalis:  
1259            Social preferences and the nature of rent seeking. *CeDEx*  
1260            *discussion paper 2008-10, University of Nottingham.*
- 1261    Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment  
1262            across societies. *Science*, 319(5868), 1362.
- 1263    Hertwig, R., & Ortmann, A. (2001). Experimental practices in  
1264            economics: A methodological challenge for psychologists?  
1265            *Behavioral and Brain Sciences*, 24(3), 383–403.
- 1266    Hilbe, C., & Traulsen, A. (2012). Emergence of responsible sanctions  
1267            without second order free riders, antisocial punishment or spite.  
1268            *Sci. Rep.*, 2.

## EXPLAINING ANTISOCIAL PUNISHMENT

- 1269 Hill, E. M., Jenkins, J., & Farmer, L. (2008). Family unpredictability,  
1270 future discounting, and risk taking. *The Journal of Socio-*  
1271 *Economics*, 37(4), 1381-1396.
- 1272 Hoff, K., Kshetramade, M., & Fehr, E. (2009). Caste and Punishment.  
1273 *Research Working papers*, 1(1), 1–42.
- 1274 Hofstede, G. (2001). *Culture's consequences: Comparing values,*  
1275 *behaviors, institutions, and organizations across nations*. Sage  
1276 Publications, Inc.
- 1277 Hopfensitz, A., & Reuben, E. (2009). The Importance of Emotions for  
1278 the Effectiveness of Social Punishment. *The Economic Journal*,  
1279 119(540), 1534–1559.
- 1280 Houser, D., & Xiao, E. (2010). Inequality seeking punishment.  
1281 *Economics Letters*, 109, 20–23.
- 1282 Inglehart, R., & Baker, W. E. (2000). Modernization, cultural change,  
1283 and the persistence of traditional values. *American Sociological*  
1284 *Review*, 65(1), 19–51.
- 1285 Jensen, K. (2010). Punishment and spite, the dark side of cooperation.  
1286 *Philosophical Transactions of the Royal Society B: Biological*  
1287 *Sciences*, 365(1553), 2635.
- 1288 Kümmerli, R., Burton-Chellew, M. N., Ross-Gillespie, A., & West, S. A.  
1289 (2010). Resistance to extreme strategies, rather than prosocial  
1290 preferences, can explain human cooperation in public goods



## EXPLAINING ANTISOCIAL PUNISHMENT

- 1291            games. *Proceedings of the National Academy of Sciences of the*  
1292            *United States of America*, 107(22), 10125–10130.
- 1293    Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on  
1294            moralistic punishment. *Evolution and Human Behavior*, 28(2),  
1295            75–84.
- 1296    Kurzban, R., & Houser, D. (2005). Experiments investigating  
1297            cooperative types in humans: A complement to evolutionary  
1298            theory and simulations. *Proceedings of the National Academy of*  
1299            *Sciences of the United States of America*, 102(5), 1803–1807.
- 1300    Kutsukake, N., & Nunn, C. L. (2006). Comparative tests of reproductive  
1301            skew in male primates: the roles of demographic factors and  
1302            incomplete control. *Behavioral Ecology and Sociobiology*, 60(5),  
1303            695–706.
- 1304    Leach, C. W., Spears, R., Branscombe, N. R., & Doosje, B. (2003).  
1305            Malicious pleasure: Schadenfreude at the suffering of another  
1306            group. *Journal of Personality and Social Psychology*, 84(5), 932–  
1307            943.
- 1308    MacLean, R. C., Fuentes-Hernandez, A., Greig, D., Hurst, L. D., &  
1309            Gudelj, I. (2010). A Mixture of “Cheats” and “Co-Operators” Can  
1310            Enable Maximal Group Benefit. *PLoS Biology*, 8(9).
- 1311    Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A.,  
1312            Cardenas, J. C., ... Tracer, D. (2008). More “altruistic”

## EXPLAINING ANTISOCIAL PUNISHMENT

- 1313 punishment in larger societies. *Proceedings of the Royal Society*  
1314 *B: Biological Sciences*, 275(1634), 587–592.
- 1315 Marlowe, F. W., Berbesque, J. C., Barrett, C., Bolyanatz, A., Gurven, M.,  
1316 & Tracer, D. (2011). The “spiteful” origins of human cooperation.  
1317 *Proceedings of the Royal Society B: Biological Sciences*, 278,  
1318 2159–2164.
- 1319 Masclet, D., Noussair, C., Tucker, S., & Villeval, M. C. (2003).  
1320 Monetary and nonmonetary punishment in the voluntary  
1321 contributions mechanism. *American Economic Review*, 93(1),  
1322 366–380.
- 1323 Maynard Smith, J. (1964). Group selection and kin selection. *Nature*,  
1324 201(4924), 1145–1147.
- 1325 Maynard Smith, J. (1974). The Theory of Games and the Evolution of  
1326 Animal Conflicts. *Journal of Theoretical Biology*, 47, 209–21.
- 1327 McLeish, K. N., & Oxoby, R. J. (2007). Identity, cooperation, and  
1328 punishment. *IZA Discussion Paper 2572*.
- 1329 Nelson-Flower, M. J., Hockey, P. A. ., O’Ryan, C., Raihani, N. J., Du  
1330 Plessis, M. A., & Ridley, A. R. (2011). Monogamous dominant  
1331 pairs monopolize reproduction in the cooperatively breeding pied  
1332 babbler. *Behavioral Ecology*, 22(3), 559.
- 1333 Nettle, D. (2006). The Evolution of Personality Variation in Humans and  
1334 Other Animals. *American Psychologist*, 61(6), 622–631.

- 1335 Nikiforakis, N. (2008). Punishment and counter-punishment in public  
1336 good games: Can we really govern ourselves? *Journal of Public*  
1337 *Economics*, 92(1-2), 91–112.
- 1338 Nikiforakis, N. (2010). Feedback, punishment and cooperation in public  
1339 good experiments. *Games and Economic Behavior*, 68(2), 689–  
1340 702.
- 1341 Ones, U., & Putterman, L. (2007). The ecology of collective action: A  
1342 public goods and sanctions experiment with controlled group  
1343 formation. *Journal of Economic Behavior & Organization*, 62(4),  
1344 495–521.
- 1345 Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and  
1346 without a sword: Self-governance is possible. *The American*  
1347 *Political Science Review*, 86(2), 404–417.
- 1348 Page, T., Putterman, L., & Garcia, B. (2008). Getting punishment right:  
1349 Do costly monitoring or redistributive punishment help? *Brown*  
1350 *University discussion paper 2008-1*.
- 1351 Patt, A., & Zeckhauser, R. (2000). Action bias and environmental  
1352 decisions. *Journal of Risk and Uncertainty*, 21(1), 45–72.
- 1353 Powers, S. T., Taylor, D. J., & Bryson, J. J. (2012). Punishment can  
1354 promote defection in group-structured populations. *Journal of*  
1355 *Theoretical Biology*, 311(0), 107–116.

## EXPLAINING ANTISOCIAL PUNISHMENT

- 1356 Price, M. E. (2005). Punitive sentiment among the Shuar and in  
1357 industrialized societies: Cross-cultural similarities. *Evolution and*  
1358 *Human Behavior*, 26(3), 279–287.
- 1359 Queller, D., C. (1994). Genetic relatedness in viscous populations.  
1360 *Evolutionary Ecology*, 8(1), 70–73.
- 1361 Rand, D. G., Armao, I. V., Joseph, J., Nakamaru, M., & Ohtsuki, H.  
1362 (2010). Anti-social punishment can prevent the co-evolution of  
1363 punishment and cooperation. *Journal of Theoretical Biology*.
- 1364 Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A.  
1365 (2009). Positive interactions promote public cooperation. *Science*,  
1366 325(5945), 1272–1275.
- 1367 Rand, D. G., & Nowak, M. A. (2011). The evolution of antisocial  
1368 punishment in optional public goods games. *Nat Commun*, 2, 434.
- 1369 Sääksvuori, L., Mappes, T., & Puurtinen, M. (2011). Costly punishment  
1370 prevails in intergroup conflict. *Proceedings of the Royal Society*  
1371 *B: Biological Sciences*.
- 1372 Saijo, T., & Nakamura, H. (1995). The spite dilemma in voluntary  
1373 contribution mechanism experiments. *Journal of Conflict*  
1374 *Resolution*, 39, 535–560.
- 1375 Santos, M. dos, Rankin, D. J., & Wedekind, C. (2011). The evolution of  
1376 punishment through reputation. *Proceedings of the Royal Society*  
1377 *B: Biological Sciences*, 278(1704), 371–377.

## EXPLAINING ANTISOCIAL PUNISHMENT

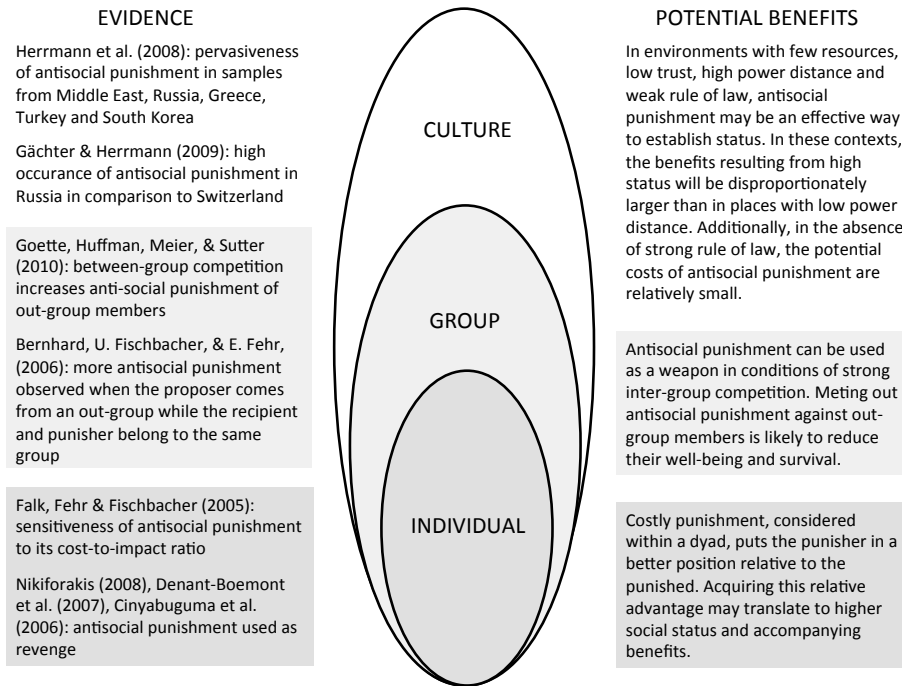
- 1378 Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary  
1379 Theory and the Ultimate–Proximate Distinction in the Human  
1380 Behavioral Sciences. *Perspectives on Psychological Science*,  
1381 6(1), 38–47. doi:10.1177/1745691610393528
- 1382 Setchell, J. M., Charpentier, 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.

## EXPLAINING ANTISOCIAL PUNISHMENT

- 1400 Thierry, B. (2005). Integrating proximate and ultimate causation: Just  
1401 one more go! *Current Science*, 89, 1180–1183.
- 1402 West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics:  
1403 altruism, cooperation, mutualism, strong reciprocity and group  
1404 selection. *Journal of Evolutionary Biology*, 20(2), 415–432.
- 1405 Wilson, D. S. (2004). What is wrong with absolute individual fitness?  
1406 *Trends in Ecology & Evolution*, 19(5), 245–248.
- 1407 Wu, J.-J., Zhang, B.-Y., Zhou, Z.-X., He, Q.-Q., Zheng, X.-D.,  
1408 Cressman, R., & Tao, Y. (2009). Costly punishment does not  
1409 always increase cooperation. *Proceedings of the National*  
1410 *Academy of Sciences*, 106(41), 17448–17451.
- 1411 Yamagishi, T. (1986). The provision of a sanctioning system as a public  
1412 good. *Journal of Personality and Social Psychology*, 51(1), 110–  
1413 116.

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

1415 with its possible benefits



1416

1417 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

1418 P4's behavior is a classical example of altruistic punishment. A cooperative individual  
1419 P4 suffers a cost to punish P2 who contributed less than the group average. As a result  
1420 of this punishment, P2's contribution increases in the next PGG round. Consider the  
1421 behavior of P1 who punished P2. As a result of the punishment, P2 increased their  
1422 contributions. Therefore, P1's punishment can be called functionally altruistic. At the  
1423 same time, this punishment would be defined as antisocial (*sensu* Herrmann et al.,  
1424 2008) because P1's original contribution, which is lower than P2's contribution, is  
1425 treated as a reference level.