Theory Dec. (2014) 76:529–545
DOI 10.1007/s11238-013-9382-3

Inequality aversion and antisocial punishment

Christian Thöni

Published online: 31 May 2013

Abstract Antisocial punishment—punishment of pro-social cooperators—has shown to be detrimental for the efficiency of informal punishment mechanisms in public goods games. The motives behind antisocial punishment acts are not yet well understood. This article shows that inequality aversion predicts antisocial punishment in public goods games with punishment. The model by Fehr and Schmidt (Q J Econ 114(3): 817–868, 1999) allows to derive conditions under which antisocial punishment occurs. With data from three studies on public goods games with punishment I evaluate the predictions. A majority of the observed antisocial punishment acts are not compatible with inequality aversion. These results suggest that the desire to equalize payoffs is not a major determinant of antisocial punishment.

Keywords Antisocial punishment · Inequality aversion · Public goods · Informal punishment · Experimental economics

JEL Classification D03 · H41 · C72 · C91

1 Introduction

A large strand of literature in experimental social psychology and experimental economics has demonstrated that informal sanctions can solve social dilemmas.[1] The upshot of this literature is that, when people are able to punish others dependent on their contributions, they do so in a way that free riding is no longer profitable and even

[1] See e.g. Yamagishi (1986); Ostrom et al. (1992) and Fehr and Gächter (2000, 2002).

C. Thöni (✉)
Centre Walras-Pareto, Quartier UNIL-Dorigny, Bâtiment Internef, University of Lausanne, 1015 Lausanne, Switzerland
e-mail: christian.thoeni@unil.ch

selfish subjects find it worthwhile to contribute. The public goods game is then no longer a social dilemma and the groups manage to avoid the tragedy of the commons (Hardin 1968). This is very remarkable, because the punishment of free riders is per construction also plagued by free rider incentives: in the best world I would have others educate the free riders to become contributors and enjoy their contributions without engaging in costly punishment myself.

Recent experimental evidence, however, puts the universality of these results into question. Gächter et al. (2005) began investigating public goods games with punishment in different cultures and found that the punishment option is hardly effective in enhancing cooperation in some of their subject pools. They presented preliminary evidence that the variation in cooperation across their subject pools is connected to the use of the punishment option. Interestingly, the differences are not in the way subjects treat free riders, but in the way they treat cooperative subjects. For punishment targeted to subjects with an equal or a higher contribution than the punisher Herrmann et al. (2008) coined the term 'antisocial punishment'.[2] They investigate 16 culturally diverse subject pools and show that there is a clear-cut connection between the prevalence of antisocial punishment and the effectiveness of the punishment option in fostering cooperation. In subject pools where antisocial punishment is frequent subjects do not profit from the punishment option and earn lower profits than without the punishment option.

Given that antisocial punishment is a major obstacle for cooperation it is important to understand the motives behind antisocial punishment. Fehr and Gächter (2000, p. 990) devote a footnote to the causes of antisocial punishment. They mention random error, improvement of the relative position (status preferences), and revenge for anticipated or past punishment. Herrmann et al. (2008) provide a more extensive account for the causes of antisocial punishment, adding 'do-gooder derogation' (Monin 2007), and a desire to punish non-conformists to the list. In the data they find evidence for the revenge explanation but their experiment is not designed to differentiate between various motives behind antisocial punishment. To date there is no experimental study that systematically explores the causes of antisocial punishment.

This paper brings an additional, maybe surprising reason for antisocial punishment into the discussion. In the next section I use the model of Fehr and Schmidt (1999) to show that inequality aversion predicts antisocial punishment in many cases. Consider a situation where (i) cooperative players are faced with a free rider and (ii) not all cooperators are willing to punish the free rider. Inequality aversion predicts that those who punish do not only punish the free rider, but also the cooperative players

[2] There are different definitions for the punishment of cooperative subjects in the literature. Herrmann et al. (2008) focus on the bilateral comparison of contributions between punisher and punishee. Falk et al. (2005) investigate motives behind the punishment decision in the prisoners' dilemma and call the punishment of cooperative subjects 'spiteful punishment', indicating that they see the motive to increase payoff differences as a determinant for such punishment acts (see also Masclet et al. 2003). Cinyabuguma et al. (2006) define 'perverse punishment' as punishing a subject who contributes more than the group average. They investigate whether second order punishment, i.e., punishing the punishers, eliminates perverse punishment. Nikiforakis (2008) addresses a similar question in a different design. Further data on antisocial punishment are reported by Anderson and Putterman (2006); Ertan et al. (2009), and Gächter and Herrmann (2009, 2011).

who do not punish. This happens for purely material reason: it ensures that the punishing players do not fall behind the players who free ride on their punishment expenditures.

Suspecting inequality aversion as a motivation for punishment acts is quite natural. For punishment of free riders Fehr and Fischbacher (2004) and Fowler et al. (2005)[3] provide evidence that egalitarian motives drive punishment decisions. In Sect. 3, I use data from three experimental studies on public goods games with punishment to investigate whether the same is true for antisocial punishment. The answer is no—a majority of antisocial punishment acts occur in situations which do not meet the conditions for antisocial punishment as explained by inequality aversion.

2 Theory

On pages 836–843 Fehr and Schmidt (1999) derive equilibria in public goods games with and without punishment. They show that in both games cooperative equilibria are possible, provided sufficiently many players are sufficiently inequality averse. In the public goods game without punishment selfish players always contribute zero. In the game with punishment even selfish players contribute as long as there is a subgroup of inequality averse players ready to punish deviators. Fehr and Schmidt's characterization of the off-equilibrium path in the game with punishment is, however, incomplete and thus fails to account for antisocial punishment. In the following I show that inequality averse players might not only punish deviant group members, but also players who contributed more than themselves.

2.1 The game

There are $n \geqslant 3$ players with an endowment of y monetary units. The game consists of two stages. In the first stage players simultaneously contribute $g_i \in [0, y]$ to a public good. After the contribution decision all players learn all contributions and their stage 1 earnings w_i, calculated as

$$w_i(\boldsymbol{g}) = y - g_i + a \sum_{j=1}^{n} g_j, \tag{1}$$

where $\boldsymbol{g}$ is the vector of all contributions and a is the marginal per capita return of the public good with $\frac{1}{n} < a < 1$. In the second stage players can punish each other bilaterally by assigning punishment points $p_{ij} \geqslant 0$, which reduce the monetary payoff of j by one unit at a cost of c units to the punisher i, and $0 < c < 1$. *Antisocial punishment* occurs when a player i punishes another player j who chose a weakly higher contribution ($g_j \geqslant g_i$). Punishment of a player j with a lower contribution ($g_j < g_i$) will be labeled as *free-rider punishment*, irrespective of the absolute level of g_j. The final monetary payoff for player i is

[3] See also Dawes et al. (2007) and Johnson et al. (2009).

$$x_i(\boldsymbol{g}, \boldsymbol{P}) = w_i(\boldsymbol{g}) - \sum_{j=1}^{n} p_{ji} - c \sum_{j=1}^{n} p_{ij}, \tag{2}$$

with $\boldsymbol{P}$ being an $n \times n$ matrix with imposed zeros on the main diagonal (masochistic punishment is usually not allowed in these experiments). Players have Fehr–Schmidt utility functions

$$u_i(\boldsymbol{x}) = x_i - \frac{\alpha_i}{n-1} \sum_{j \neq i} \max\left[x_j - x_i, 0\right] - \frac{\beta_i}{n-1} \sum_{j \neq i} \max\left[x_i - x_j, 0\right], \tag{3}$$

where every unit of payoff difference in bilateral payoff comparisons reduces utility by $\alpha_i/(n-1)$ in case the other player earns more and $\beta_i/(n-1)$ in case the other player earns less than player i. Fehr and Schmidt put restrictions on the preference parameters: both kinds of inequality hurt or are neutral ($\alpha_i, \beta_i \geqslant 0$), having less than others hurts weakly more than having more ($\alpha_i \geqslant \beta_i$), and having more than others must not hurt so badly that an individual would want to burn his own money to restore equality ($\beta_i < 1$).

Fehr and Schmidt (1999, p. 841) show that, if sufficiently many players are inequality averse, subgame perfect equilibria with non-minimal contributions $g_i = g \in (0, y]\ \forall i$ exist. In these equilibria also selfish players contribute g, because they face a credible threat of punishment by the inequality averse players. In equilibrium no punishment occurs. To explain any kind of punishment, we need to assume that at least one player deviates from the equilibrium strategy and chooses $g_i < g$. In the following we restrict our attention to the subgame where p_{ij} is chosen; I refer to this subgame as the 'punishment subgame'.[4] I present three simple numerical examples to illustrate how players with Fehr–Schmidt preferences punish other players. A general solution of the punishment subgame is provided in the Appendix. The first two examples focus on a situation in which only one player is willing to punish other players.

2.2 The lone enforcer

Example 1 Consider the game with $\{n, y, a, c\} = \left\{3, 20, \frac{1}{2}, \frac{1}{3}\right\}$. Assume the group consists of two players, 1 and 2, with selfish preferences, i.e. $\{\alpha_i, \beta_i\} = \{0, 0\}$, $i = 1, 2$, while the third player E shall be inequality averse with $\{\alpha_E, \beta_E\} = \{4, .6\}$, henceforth called the enforcer.

Consider a situation where E and player 1 contributed 20, while player 2 contributed 0.[5] The first part of Table 1 shows the utilities that result from this situation if no punishment is exerted. How should players punish each other? The two selfish

[4] In fact, the analysis that follows is applicable to any situation where members of a group of n players, endowed with some income $\boldsymbol{w}$, can mutually reduce their incomes by the punishment mechanism described in Eq. (2).

[5] Fehr and Schmidt (1999, p. 841) show that the composition of types and parameters assumed here allows for a fully cooperative equilibrium. To explain punishment we need to assume that one player deviates in the contribution stage.

Table 1 Antisocial punishment in a three player public goods game with punishment

Player	g_i	No punish.		Fehr–Schmidt Prop. 5			Optimal		
		w_i	u_i	p_{Ej}	x_i	u_i	p_{Ej}	x_i	u_i
E	20	20	**−20**		10	**−10**		0	**0**
1	20	20	20		20	20	**20**	0	0
2	0	40	40	**30**	10	10	**40**	0	0

Bold numbers indicate utilities or punishment acts of player E, the enforcer

players would certainly not punish because punishment is costly and yields no benefit to them. Player E, however, might want to punish because she does not like the fact that player 2 earns more than she does. Her utility without punishment amounts to -20, because she suffers from disadvantageous inequality towards player 2 of 20 monetary units. According to the Fehr–Schmidt utility function this difference is weighted by $\alpha_E/2$ and subtracted from her monetary income. What relief can E give herself by punishing others? According to Fehr and Schmidt's Proposition 5, E should reduce player 2's income by $p_{E2} = 30$ units (see intermediate columns in Table 1). This would equalize player 2's and E's monetary income and increase E's utility to -10.

However, this is not the equilibrium of the punishment subgame. In fact, E is not playing best response if she just punishes the deviator (player 2). The last three columns of Table 1 show the optimal punishment strategy. E maximizes her utility if she punishes *both* other players. She does so in a way to equalize all three monetary payoffs, which allows her to enjoy a utility of zero. Thus, the Fehr–Schmidt utility function does indeed predict antisocial punishment, i.e. punishment of players who are equally or more cooperative than the punisher.

Table 1 shows also that the punishment for the free rider becomes harsher once we take antisocial punishment into account. For the existence of an equilibrium with non-minimal contributions punishment has to be sufficiently strong to make deviations unprofitable. This is clearly the case in the example: had player 2 contributed 20 there would have been no punishment and all players would have earned 30. With the deviation player 2 earns 10 (according to Fehr–Schmidt) or 0 for optimal punishment. Thus, if the parameter constellation derived in Fehr–Schmidt Prop. 5 allows for a cooperative equilibrium it will also do so when we take the possibility of antisocial punishment into account.[6]

Example 2 In a next step let us consider a richer example. Table 2 depicts the situation of a public goods game with $\{n, y, a, c\} = \left\{4, 20, \frac{2}{5}, \frac{1}{3}\right\}$. To make things interesting I assume that the enforcer chose a contribution of 9, and the other players in the group chose 0, 11 and 20. This produces a set of stage 1 earnings as depicted in the third column of Table 2. Let us look at two types of enforcers, a 'weakly inequality averse' player with $\{\alpha_E, \beta_E\} = \{1, .6\}$, and a 'strongly inequality averse' player with

[6] The inclusion of antisocial punishment does not affect the general structure of the equilibria described in Fehr–Schmidt's Prop. 5, where all players contribute $g \in (0, y]$ in the first stage and punishment occurs only off equilibrium.

Table 2 Optimal punishment for a weakly or strongly inequality averse enforcer in a four player public goods game with punishment

Player	g_i	No punish.		Weakly ineq. averse			Strongly ineq. averse		
		w_i	u_i	p_{Ej}	x_i	u_i	p_{Ej}	x_i	u_i
$E_{\alpha_E=1}$ / $E_{\alpha_E=4}$	9	27	**21.4** / **12.4**		25	**21.5** / **16.5**		20	**19.2** / **19.2**
1	0	36	36	**6**	30	30	**16**	20	20
2	11	25	25		25	25	**5**	20	20
3	20	16	16		16	16		16	16

$\{\alpha_E, \beta_E\} = \{4, .6\}$. These two types will have different optimal punishment patterns. The three columns in the middle of Table 2 show the optimal punishment strategy of the weakly inequality averse player. Such a player maximizes his utility by punishing the richest among the other players (the free rider). However, punishment is only utility maximizing as long as E's income does not drop below another player's income. Consequently, as soon as player E's income is as low as the next lower income (in this case player 2's income), he does not exert more punishment, despite the fact that player 1 still earns more than he does.

A strongly inequality averse player does not stop here. Her much higher α_E means that the income difference towards player 1 still weighs heavily on her well-being. A strongly inequality averse player prefers to further decrease player 1's income. However, in order to avoid falling behind player 2, she punishes both other players with weakly higher incomes. Player E's utility is maximized when the incomes of player 1 and 2 are equal to her own income.

What does this mean for antisocial punishment? The weakly inequality averse player in our example does not engage in antisocial punishment, whereas the strongly inequality averse player does mete out punishment to a player who was more cooperative than herself. This example demonstrates the necessary conditions for antisocial punishment: (i) player E must be sufficiently inequality averse to be willing to punish more than one other player, (ii) there must be at least one player with a strictly higher income than E and (iii) in the process of punishing the strictly richer player, E's income must undershoot one of the other (weakly poorer) player's incomes.

2.3 Teaming up

Example 3 What if more than one player is inequality averse enough to punish other players? Let there be a subgroup of $n' = 2$ homogeneous enforcers[7] and $\{n, y, a, c\} = \left\{4, 20, \frac{2}{5}, \frac{1}{3}\right\}$. Table 3 shows the situation of two enforcers with $\{\alpha_E, \beta_E\} = \{4, .6\}$ facing one cooperative player and one free rider. According to Proposition 5 in Fehr and Schmidt the two enforcers should punish player 2 (the free rider) such that this player's income is equal to their income. However, similar to the case of the lone

[7] To keep things simple I assume that all enforcers are homogeneous both in their preferences (equal α_E and β_E) and in their stage 1 earnings w_{Ei}.

Table 3 Optimal punishment in a four player public goods game with punishment with two enforcers

Player	g_i	No punish.		Fehr–Schmidt Prop. 5				Optimal punishment			
		w_i	u_i	p_{E1j}	p_{E2j}	x_i	u_i	p_{E1j}	p_{E2j}	x_i	u_i
$E1$	20	24	**−16**			20	**12**			19	**19**
$E2$	20	24	**−16**			20	**12**			19	**19**
1	20	24	24			24	24	**2.5**	**2.5**	19	19
2	0	44	44	**12**	**12**	20	20	**12.5**	**12.5**	19	19

enforcer, the two enforcers can improve their situation if they also punish player 1. In the example they maximize their utility by equalizing the monetary payoff of all four players. However, unlike in the case of the lone enforcer, only punishing the deviant player is also an equilibrium of the punishment subgame. If $E1$ punishes only player 2 by 12 units, then $E2$'s best response is to punish only player 2 by 12 units as well and vice versa. However, the equilibrium including antisocial punishment towards the fully contributing non-enforcer is clearly more efficient for the two enforcers.

2.4 Conditions for antisocial punishment

The three examples show that antisocial punishment motivated by Fehr–Schmidt inequality aversion happens only in order to prevent falling behind other players who do not punish.[8] The Appendix provides a general solution to the punishment subgame. A decisive prerequisite is the existence of other players who do not engage in punishment. Thus, the decision whether to engage in antisocial punishment or not depends on players' beliefs about other players' punishment behavior. This dependency on beliefs makes it difficult to identify *sufficient* conditions for antisocial punishment and an empirical investigation would require information about subjects' beliefs about all other punishments. However, for an empirical test of the predictions it is easy to derive two *necessary* conditions for the occurrence of antisocial punishment in public goods games with punishment, which hold for any belief about other player's α and β. If antisocial punishment is motivated by Fehr–Schmidt like inequality aversion, then it must hold that:

C1: A maximum earner never punishes, i.e. antisocial punishment can only occur if the punisher's stage 1 income is below the maximum stage 1 income in the group.

C2: The only reason for an enforcer to punish a weakly poorer player is to avoid falling behind this player while punishing free riders. Consequently, antisocial punishment can only occur in combination with punishment of a richer player.[9]

[8] Note that the ERC model by Bolton and Ockenfels (2000) is also capable in predicting antisocial punishment, but does so in a rather trivial way. In this model players care about their share of the total pie (and their own income). If they earn too low a share they can increase their share by punishing *any* other group member, not just the free rider. Consequently, the ERC model has no predictive power whatsoever about the direction of punishment.

[9] This holds for symmetric equilibria of the punishment subgame, i.e. equilibria where all enforcers punish the other players equally. Because the enforcers do only care about final payoffs they could freely reallocate

Proofs are provided in the Appendix. In the next step I confront these two conditions with data on antisocial punishment.

3 Empirics

In this Section I use the data of three studies on public goods games with punishment. First I look at the study by Fehr and Gächter (2002) who report data from 120 subjects playing six rounds of the public goods game with punishment with $\{n, y, a, c\} = \left\{4, 20, \frac{2}{5}, \frac{1}{3}\right\}$.[10] They use a perfect stranger matching protocol, which ensures that, during the six rounds, a subject will not meet another subject in the session more than once. Thus, from a game-theoretic perspective we observe the subjects playing six consecutive one-shot games.

Before addressing the necessary conditions for antisocial punishment I investigate the frequency of the prototypical situations for observing antisocial punishment. Recall that the situations typically require the presence of a free rider and non-punishing contributors. The overall frequency of free-rider punishment (strictly positive amounts of punishment targeted to subjects with a lower contribution) in the sample is 46.9 %, i.e. in slightly more than half of the cases subjects do not punish others with a lower contribution.[11]

In a next step I investigate whether observed acts of antisocial punishment are in line with the predictions of the Fehr–Schmidt model. The first result addresses C1.

Result 1 : Maximum earners punish

Support According to C1 all antisocial punishment must come from subjects who earn less than the maximum stage 1 income in a group. In Fehr and Gächter's data top earners punish weakly poorer subjects by 1.60 punishment points on average.[12] Antisocial punishment meted out by top earners is obviously not zero, but is that number small or large? Subjects who are not among the richest in their group mete

Footnote 9 continued
punishment points among themselves, as long as the total amount of punishment meted out by each punisher and received by each punishee remains constant. In case the group of enforcers is sufficiently large there can be asymmetric equilibria where some enforcers mete out exclusively antisocial punishment.

[10] I use only the data from the sessions with the sequence punishment—no punishment.

[11] This number holds for all situations. It might be more interesting to consider only situations similar to those described in Table 1, where there is a clear free rider. Consider the cases where three subjects contribute strictly more than the group average and one subject contributes strictly less than the group average. In 40.3 % of these cases all three high contributors punish the free rider. Consequently, in 59.7 % of the situations not all three high contributors engage in free-rider punishment, leaving scope for antisocial punishment as explained by inequality aversion.

[12] The number corresponds to the parameter p_{ij} in equation 2, i.e. to the average reduction of the punishee's income. In the experiment subjects choose 'deduction points', which reduce the punishee's income by three units and cost the punisher one unit. Deduction points are limited to integers and up to ten points per punishment act. Thus, in the experiment the optimal punishment solution presented in Table 1 is not feasible, because player 2's income cannot be reduced by more than 30 units.

out .43 punishment points to weakly poorer subjects.[13] Thus, not only do the richest subjects in the experiments by Fehr and Gächter (2002) punish other weakly poorer players—they do so even stronger than subjects who are not among the richest in the group. This difference is significant by a two-sample Wilcoxon rank-sum test ($z = -2.61,\ p = 0.008$, two-sided exact p-value, test based on independent session averages).

Instead of just focusing on whether a subject is richest or not it is also possible to consider a more fine grained measure for a subject's position in the income hierarchy before punishment. If we assign rank one to the richest subject(s) and subsequent ranks for poorer subjects (standard competition ranking) then antisocial punishment is decreasing in ranks starting with 1.60 for the richest to .48 (.32) for the second (third) richest. The poorest cannot punish antisocially by definition. Interestingly, free-rider punishment is also decreasing in ranks. The second (third, fourth) richest mete out 4.48 (3.34, 2.25) units of free-rider punishment. Thus strongest punishment of free riders does not stem from the highest contributors but from the intermediate contributors. This fits very nicely to the concept of inequality aversion because it suggests that intermediate contributors are willing to pay a larger part of the punishment costs than the poorer high contributors in their group. Thus, while free-rider punishment seems to be accessible to an inequality aversion explanation, antisocial punishment is not.

Result 2 : Subjects frequently mete out only antisocial punishment

Support According to C2 every act of antisocial punishment should be accompanied by punishment of a subject with a lower contribution than the punisher. The upper part of Table 4 classifies all 720 individual punishment vectors. In about 43 % of the cases no subject gets punished. In about 42 % of the cases only free riders (i.e. subjects with a lower contribution than the punisher) receive punishment. About 15 % (105) of the cases involve antisocial punishment. In 75 % of these cases antisocial punishment is not accompanied by free-rider punishment. Thus, contrary to the theoretical prediction, antisocial punishment is most frequently meted out without accompanying free-rider punishment.

The evidence presented clearly indicates that the two necessary conditions for antisocial punishment as explained by the Fehr–Schmidt model are very frequently not met for actual acts of antisocial punishment. The analysis could, however, be influenced by the fact that subjects in the experiments by Fehr and Gächter (2002) play six consecutive games. Despite the use of the perfect stranger matching we cannot be sure that the game reflects a true one-shot game. The fact that the authors observe a time trend in average contributions even suggests that the six rounds are not identical. A recent study by Gächter and Herrmann (2009) provides data from true one-shot games with $\{n, y, a, c\} = \left\{3, 20, \frac{1}{2}, \frac{1}{3}\right\}$. In a next step I apply the same analysis to this

[13] This comparison controls for the fact that a subject who is among the richest players has more 'occasions' to engage in antisocial punishment than a poorer subject, because it is the *average* number of punishment points assigned in all bilateral comparisons with weakly poorer subjects. Looking at absolute numbers the difference becomes even stronger. From a total of 1188 units of payoff reduction by antisocial punishment 921 (78 %) are caused by the richest subjects in the group.

Table 4 No. of cases with no, antisocial, free rider, or both kinds of punishment, with row and column totals

	Free-rider punishment		
	No	Yes	
Fehr and Gächter (2002)			
Antisocial punishment			
No	310	305	615
Yes	79	26	105
	389	331	720
Gächter and Herrmann (2009)			
Antisocial punishment			
No	130	75	205
Yes	43	17	60
	173	92	265

Data source upper part Fehr and Gächter (2002), lower part: Gächter and Herrmann (2009); own calculation

dataset. The data contain observations from 265 subjects in four locations in Russia and Switzerland.[14] For the moment I ignore the fact that motivations for antisocial punishment might differ across subject pools. This point will be addressed below.

The results from the first and second dataset are qualitatively very similar. In the data by Gächter and Herrmann (2009) average antisocial punishment meted out by top earners is 2.24 compared to .95 by all other subjects. The difference is significant ($z = -3.04$, $p = 0.002$, test based on individual observations). The lower part of Table 4 also shows that in this dataset the majority of antisocial punishment acts (72 %) is not accompanied by free-rider punishment.

Cheung (2012) provides further evidence on one-shot games using the same parameters as Gächter and Herrmann, but with the punishment decision elicited by the strategy method. Subjects choose punishments for all possible combinations of other subjects' contributions. This allows for a much more detailed analysis of punishment behavior.[15] The paper does not provide exact numbers on C1 but from the figures it becomes clear that there is a considerable amount of punishment meted out by the maximum earners. For C2 Cheung (2012, p. 24) provides a detailed analysis showing that 21 out of 280 cases potentially fit to inequality aversion. However, the majority of the subjects involved does also punish antisocially when being among the maximum earners, which is clearly incompatible with the prediction.

Herrmann et al. (2008) report data from 16 subject pools from various cultural backgrounds. They observe punishment decisions from experiments with ten times repeated public goods game with punishment and $\{n, y, a, c\} = \left\{4, 20, \frac{2}{5}, \frac{1}{3}\right\}$. Their data contain observations from 1,120 subjects in 16 cities around the globe. The 16 subject pools differ markedly in the frequency and strength of antisocial punishment. Thus, even if inequality aversion fails to explain antisocial punishment in the three

[14] Like in the first dataset I use only the data from the sessions which started with the one-shot public goods game with punishment.

[15] The data does, however, still not allow to identify sufficient conditions for antisocial punishment. To do so would require control over the subjects' beliefs about all other punishment decisions.

Table 5 Patterns of antisocial punishment across subject pools

	Antisocial punishment when…		Perc. only anti-social punishment
	Richest	Not richest	
Athens	4.03	0.86	86.4 %
Bonn	0.34	0.48	68.3 %
Boston	0.26	0.15	93.3 %
Chengdu	0.50	0.34	89.4 %
Copenhagen	0.33	0.28	84.2 %
Dnipropetrovs'k	1.37	0.65	80.9 %
Istanbul	1.32	0.72	72.4 %
Melbourne	0.73	0.32	80.0 %
Minsk	1.44	1.39	64.5 %
Muscat	3.81	3.23	67.6 %
Nottingham	0.37	0.13	83.3 %
Riyadh	1.86	2.23	57.7 %
Samara	2.34	1.02	73.7 %
Seoul	0.60	0.40	79.0 %
St.Gallen	0.49	0.34	81.3 %
Zurich	0.48	0.20	85.1 %
Total	1.03	0.79	75.5 %

Data source Herrmann et al. (2008), own calculation

datasets discussed so far it is possible that this explanation works well in some of the subject pools observed by Herrmann et al. (2008).[16]

Table 5 shows the two measures discussed above for each of the 16 subject pools separately. The first column shows the average punishment meted out by subjects who are maximum earners in the specific period. This is the number that should be zero according to the prediction. Obviously there is a huge variation in the strength of antisocial punishment. Average punishment differs by more than a factor ten when comparing the extreme cases. The second column shows the average antisocial punishment by subjects who are not among the richest in their group. In 14 out of the 16 subject pools subjects choose more antisocial punishment when they are top earners compared to when there are others in the group with higher stage 1 incomes.

The third column in Table 5 shows the fraction of antisocial punishment acts not accompanied by free-rider punishment. Everywhere, in at least 50 % of the cases with antisocial punishment subjects mete out *only* antisocial punishment. To conclude,

[16] The analysis presented here uses data from a dynamic game to test a static prediction. This could be problematic because of strategic incentives in early rounds of the game. To check whether the results are robust with regard to this concern I ran the analysis for the last period only. The last punishment subgame played presents a true one-shot game. The results remain qualitatively unchanged.

despite the huge differences in the level of antisocial punishment across subject pools the pattern of antisocial punishment acts does not differ substantially. In none of the subject pools antisocial punishment seems to be motivated by inequality aversion.

4 Conclusion

To test whether observed acts of antisocial punishment can be explained by inequality aversion as formalized by Fehr and Schmidt (1999) I derived two conditions: (i) the punishing subject must not be among the richest in the group and (ii) antisocial punishment requires accompanying acts of free-rider punishment. Empirical evidence on antisocial punishment shows that a majority of the punishment acts occur in situations in which these two criteria are not met. Thus, inequality aversion is not a key determinant of antisocial punishment. This paper contributes to the literature on motives behind the punishment acts. Fowler et al. (2005) and Fehr and Fischbacher (2004) argue that egalitarian motives drive punishment decisions. This might be the case for free-rider punishment but most likely not for antisocial punishment. The results presented here are in line with Falk et al. (2005), who compare the frequently used one-to-three punishment technology to a one-to-one punishment technology ($c = 1$). In the latter regime no Fehr–Schmidt player would use punishment. However, in their experiments with Prisoners' Dilemma games cooperators punished defectors also in the treatments with $c = 1$, despite the fact that payoff differences could not be reduced.[17] Defectors' punishment of cooperators was observed if $c = \frac{1}{3}$ but vanished in the treatments with the one-to-one punishment, indicating that relative payoff concerns may be an important determinant of antisocial punishment.[18] However, the relative payoff maximization hypothesis has a serious flaw when applied to public goods experiments with four players and one-to-three punishment mechanism (like used e.g. in Fehr and Gächter 2002 or Herrmann et al. 2008): Given these parameters no player can improve her relative position towards the others by punishment. Due to $c = \frac{1}{3}$ every unit invested in the punishment of another player j increases the relative position towards j by two units (j loses three, punisher loses one). However, at the same time the relative position towards the other two players decreases by one unit. Thus, if a relative payoff maximizer weighs all bilateral comparison equally, then punishment does not improve the relative position at all.[19]

Understanding the determinants of antisocial punishment is crucial, given that the efficiency of sanctioning mechanisms strongly depends on the frequency and strength of antisocial punishment, leading to strikingly different macro results across different cultures (Gächter et al. 2010). The contribution of this paper is to show that interdependent preferences as formalized in the Fehr–Schmidt model provides a rationale for

[17] See also Masclet and Villeval (2008) and Egas and Riedl (2008) for data on one-to-one punishment.

[18] See also Houser and Xiao (2010).

[19] Such a relative payoff maximizer could be characterized by having a utility function as shown in Eq. (3) with $\beta < 0$. For example, a player with $\beta = -\alpha$ would always want to punish other players (irrespective of whether they are poorer or richer) if $cn < 1 + c$ and $\alpha > \frac{c(n-1)}{1+c-cn}$. For the parameters used in Fehr and Gächter (2002) and Herrmann et al. (2008) the first condition holds with equality and requirements for α go to infinity.

antisocial punishment. Interestingly, antisocial punishment as rationalized by Fehr–Schmidt preferences is kind of a higher order punishment, targeted to players who are unwilling to bear their share of the punishment costs necessary to discipline the free riders (see e.g. Denant-Boemont et al. 2007). Seen from this angle, antisocial punishment might not be that 'perverse' after all.

The data, however, suggest that other forces must be at work. A likely candidate is revenge. Subjects might engage in antisocial punishment to take revenge for punishment received in past periods or for anticipated punishment in the current period. Herrmann et al. (2008) provide evidence for this, showing that antisocial punishment is stronger when a subject was punished in the previous period. A second likely candidate is that some subjects might simply enjoy to destroy others' property. There is experimental evidence that, in some settings, subjects reduce other subjects' incomes, even if this comes at a cost to themselves (see e.g. Abbink and Herrmann 2011; Abbink and Sadrieh 2009; Zizzo 2003; Zizzo and Oswald 2001). For public goods games with punishment under uncertainty Grechenig et al. (2010) show that, contrary to the expectation, being in doubt about other subjects' actions does not discourage from using the punishment option. Thus, a pure 'appetite for destruction' might account for some of the antisocial punishment acts. This would be compatible with the observation that antisocial punishment is, to a large extent, meted out by subjects with a high payoff. These are the subjects who can afford to spend money on their pleasure to reduce the income of others without falling behind others. An experimental analysis that cleanly separates these causes is yet to be conducted.

Acknowledgments For helpful comments and suggestions I thank two anonymous referees, Simon Gächter, Louis Putterman, Jonathan Schulz and the participants of the Thurgau Experimental Economics Meeting and the meeting of the Economic Science Association.

Appendix

Prerequisites

There are two types of players: a homogenous subgroup of $n' < n$ enforcers $E1$, $E2$, …, En' with $\alpha, \beta > 0$ and $n - n'$ other players with $\alpha, \beta = 0$. All enforcers have identical provisional income, i.e. $w_{E1} = w_{E2} = \ldots = w_{En'}$. In case of $n' > 1$ there is usually an infinite number of equilibria in which all enforcers spend an equal amount on punishment. In the following I will derive the symmetric equilibria most efficient for the enforcers. The homogeneity assumption facilitates the derivation of equilibria because all enforcers either punish or do not punish. In equilibrium all enforcers punish equally and no inequality among them arises.

What are the conditions under which Fehr–Schmidt players might engage in antisocial punishment? Enforcers face a distribution of provisional payoffs $\boldsymbol{w}$ and seek to maximize their utility by choosing a punishment vector $\boldsymbol{p_{Ek}}$ for all $k = 1, \ldots, n'$. In the group there are $n - n'$ other players. These players shall be ordered according their initial income w_i, such that the incomes of the players are $w_1 \geqslant w_2 \geqslant \ldots \geqslant w_r \geqslant w_{Ek} > w_{r+1} \geqslant \ldots \geqslant w_{n-n'}$, i.e. Player 1 is the richest player, r players are weakly richer than the enforcers, and $n - n' - r$ players are strictly poorer. The variable r is an

indicator for the enforcers' position in the income hierarchy. For $r = 0$ the enforcers are the strictly richest players and there are $n - n'$ poorer players. In case of $r = n - n'$ the enforcers are (among) the poorest players.

Proof of C1 Will the enforcers ever punish players with a lower income? Due to $c < 1$ this increases inequality towards the punished player. Furthermore, it reduces the enforcers' incomes by c and increases the inequality towards all players who are or become richer than the enforcers. The only benefit the enforcers can get from punishing a poorer player is that the reduction of their income reduces inequality towards the other players with lower incomes by c. The case most favorable for punishing a poorer player is thus the situation where a single enforcer is the richest player, i.e. $r = 0$. In such a situation it would be utility enhancing to punish another player if benefits outweigh costs, i.e.

$$\frac{\beta}{n-1}(n-2)c > c + \frac{\beta}{n-1}(1-c). \tag{4}$$

Rearranging leads to $\beta - \frac{\beta}{c(n-1)} > 1$. This inequality can only be satisfied for $\beta > 1$, which is ruled out by the parameter restrictions of the Fehr–Schmidt model. Thus, irrespective of the position in the income distribution the enforcers will never punish a player with lower income than themselves. Note that this does not exclude that punishment of players with $w_i < w_E$ eventually takes place, when the enforcers become poorer than other players due to their punishment of free riders. However, if the enforcers are among the richest players in the first place, then this situation cannot occur and, consequently, antisocial punishment can be ruled out. □

Proof of C2 What is the structure of the optimal punishment strategy? The examples in the main text already demonstrated that it can be utility enhancing for the enforcers to punish free riders. The crucial question is not whether, but *how many* richer players the enforcers are ready to punish.[20] Punishing r weakly richer player has costs and benefits: (i) punishment has direct costs of rc, (ii) due to $c < 1$ this decreases the disadvantageous inequality towards the punished player and (iii) it reduces the enforcers' payoff advantage towards the other $n - r - n'$ players with lower income by c. Taken together, punishing all r weakly richer players pays if [21]

$$\underbrace{rc}_{\text{i}} < \underbrace{\frac{\alpha}{n-1}r(n'-rc)}_{\text{ii}} + \underbrace{\frac{\beta}{n-1}(n-r-n')rc}_{\text{iii}}. \tag{5}$$

This expression allows to identify the effect of changes in the preference parameters: punishing richer players is more likely to be profitable if the enforcers' inequality

[20] Due to the linearity in payoff differences in the Fehr–Schmidt utility function E is indifferent between shifting punishment points from one richer player to another richer player. Thus, if a player E is ready to punish one richer player by, say ϵ, then she is also ready to punish two richer players by $\frac{\epsilon}{2}$.

[21] Here I assume that all enforcers use the same punishment strategy so that no inequality towards other enforcers arises. In case of $n' > 1$ Eq. 5 describes a joint optimization for all enforcers.

aversion becomes stronger (α and β). Equalizing benefits and cost in Eq. (5) and solving for r gives the maximum integer number of other players that will be punished

$$\tilde{r} = \left\lfloor \frac{\alpha n'}{c\,(\alpha+\beta)} + \frac{\beta n - \beta n' - n + 1}{(\alpha+\beta)} \right\rfloor. \tag{6}$$

It is easy to show that $\tilde{r}$ increases in the number of enforcers n'. The expression is decreasing in n, which is due to the fact that in larger groups the inequality towards the r richer players has less weight in an enforcer's utility function. Furthermore, $\tilde{r}$ is decreasing in c, i.e. more expensive punishment reduces the number of other players the enforcers are willing to punish.

What is the optimal amount of punishment? As demonstrated in Table 2 we have to check whether punishment is constrained by the income of the next poorer player or not. For *unconstrained* punishment all enforcers and weakly richer players have the same final payoff, i.e. we have to solve the following system of equations:

$$w_{Ek} - c\sum_{j=1}^{r} p_{Ekj} = w_i - \sum_{k=1}^{n'} p_{Eki} \qquad \forall\, i = 1,\ldots,r \quad \text{and} \quad k = 1,\ldots,n', \tag{7}$$

where the final income of an enforcer k is on the left hand side and the right hand side shows the income of a weakly richer player i. To simplify matters I assume that punishment of a player i is split equally among the enforcers. This allows to replace p_{Ekj} by p_{Ej} and the sum on the right hand side by $n' p_{Ei}$. Thus, dependent on r, optimal punishment points are

$$p_{Ei} = \frac{w_i}{n'} + \frac{c\sum_{j=1}^{r} w_j - n' w_E}{n'(n' - rc)}. \tag{8}$$

Total expenditures of an enforcer for punishment in the unconstrained case are, therefore,

$$\pi_E^r = c\sum_{i=1}^{r} p_{Ei} = \frac{c\sum_{i=1}^{r} w_i - rcw_E}{n' - rc}. \tag{9}$$

Clearly the amount of punishment necessary to bring down the r weakly richer players decreases in n'. When is punishment unconstrained? If $w_E - \pi_E^r \geqslant w_{r+1}$ then there is enough 'room' to punish all richer players without undershooting the income of the next poorer player.

Otherwise the enforcers are in the *constrained* case. Here the enforcers' incomes will touch w_{r+1} before they could equate all incomes of the r richer players with their own incomes. If this happens the number of weakly richer players increases by one and the optimal punishment for $r+1$ comes into action. The enforcers will include further players into the group of punishees until either (i) there is no strictly richer player anymore or (ii) the group of weakly richer players exceeds $\tilde{r}$. To conclude, punishment expenditures depend on the distribution of the preliminary incomes $\boldsymbol{w}$ which

is characterized by r, the number of weakly richer players. Punishment expenditures are

$$\pi_E(\boldsymbol{w}) = \begin{cases} 0 & \text{if } r = 0 \text{ or } r > \tilde{r} \\ \pi_E^r & \text{else if } \pi_E^r \leqslant w_E - w_{r+1} \\ \pi_E^{r+1} & \text{else if } \pi_E^{r+1} \leqslant w_E - w_{r+2} \\ \vdots & \\ \pi_E^{\tilde{r}} & \text{else if } \pi_E^{\tilde{r}} \leqslant w_E - w_{\tilde{r}+1} \text{ or } r = \tilde{r} \\ w_E - w_{\tilde{r}+1} & \text{else} \end{cases} \tag{10}$$

In all but the first and last case the enforcers reduce the incomes of richer players such that all r players earn the same income as the enforcers. In doing so the group of weakly richer players might increase up to a maximum of $\tilde{r}$. Depending on the provisional income of player $\tilde{r}+1$ the enforcers equalize the incomes of all $\tilde{r}$ players or mete out punishment such that their income is equal to $w_{\tilde{r}+1}$. In the latter case the free riders will keep some of their monetary payoff advantage relative to the enforcers (as demonstrated by the case of the weakly inequality averse player in Table 2).

To conclude, if none of the players are richer than the enforcers then no one will be punished. If some of the players are richer and others are poorer than the enforcers, then the poorer players are punished if and only if the enforcers *become* poorer than some of these players due to the punishment of free riders. Consequently, antisocial punishment can only occur *in combination* with free-rider punishment. □

References

Abbink, K., & Herrmann, B. (2011). The moral costs of nastiness. *Economic Inquiry*, *49*(2), 631–633. doi:10.1111/j.1465-7295.2010.00309.x.

Abbink, K., & Sadrieh, A. (2009). The pleasure of being nasty. *Economics Letters*, *105*, 306–308. doi:10.1016/j.econlet.2009.08.024.

Anderson, C. M., & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, *54*(1), 1–24. doi:10.1016/j.geb.2004.08.007.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193. doi:10.1257/aer.90.1.166.

Cheung, S. L. (2012). *New Insights into Conditional Cooperation and Punishment from a Strategy Method Experiment*. Economics Working Paper Series 2012–1, University of Sydney.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, *9*(3), 265–279. doi:10.1007/s10683-006-9127-z.

Dawes, C. T., Fowler, J. H., McElreath, R., Johnson, T., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, *446*(7137), 794–796. doi:10.1038/nature05651.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, *33*(1), 145–167. doi:10.1007/s00199-007-0212-0.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1637), 871–878. doi:10.1098/rspb.2007.1558.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, *53*(5), 495–511. doi:10.1016/j.euroecorev.2008.09.007.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *73*(6), 2017–2030. doi:10.1111/j.1468-0262.2005.00644.x.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. doi:10.1016/S1090-5138(04)00005-4.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994. doi:10.1257/aer.90.4.980.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140. doi:10.1038/415137a.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*(3), 817–868. doi:10.1162/003355399556151.

Fowler, J. H., Johnson, T., & Smirnov, O. (2005). Egalitarian motive and altruistic punishment. *Nature*, *433*(7021), E1–E1. doi:10.1038/nature03256.

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 791–806. doi:10.1098/rstb.2008.0275.

Gächter, S., & Herrmann, B. (2011). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review*, *55*(2), 193–210. doi:10.1016/j.euroecorev.2010.04.003.

Gächter, S., Herrmann, B., & Thöni, C. (2005). Cross-cultural differences in norm enforcement. *Behavioral and Brain Sciences*, *28*(6), 822–823. doi:10.1017/S0140525X05290143.

Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*, 2651–2661. doi:10.1098/rstb.2010.0135.

Grechenig, K., Nicklisch, A., & Thöni, C. (2010). Punishment despite reasonable doubt—A public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies*, *7*(4), 847–867. doi:10.1111/j.1740-1461.2010.01197.x.

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*(3859), 1243–1248. doi:10.1126/science.162.3859.1243.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367. doi:10.1126/science.1153808.

Houser, D., & Xiao, E. (2010). Inequality-seeking punishment. *Economics Letters*, *109*(1), 20–23. doi:10.1016/j.econlet.2010.07.008.

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters*, *102*(3), 192–194. doi:10.1016/j.econlet.2009.01.003.

Masclet, D., Noussair, C. N., Tucker, S., & Villeval, M.-C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review*, *93*(1), 366–380. doi:10.1257/000282803321455359.

Masclet, D., & Villeval, M.-C. (2008). Punishment, inequality, and welfare: A public good experiment. *Social Choice and Welfare*, *31*(3), 475–502. doi:10.1007/s00355-007-0291-7.

Monin, B. (2007). Holier than me? Threatening social comparison in the moral domain. *International Review of Social Psychology*, *20*(1), 53–68.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, *92*, 91–112. doi:10.1016/j.jpubeco.2007.04.008.

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

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

Zizzo, D. J. (2003). Money burning and rank egalitarianism with random dictators. *Economics Letters*, *81*, 263–266. doi:10.1016/S0165-1765(03)00190-3.

Zizzo, D. J., & Oswald, A. J. (2001). Are people willing to pay to reduce others' incomes? *Annales d'Economie et de Statistique*, *63–64*, 39–65.