Punishment Can Support Cooperation Even When Punishable

Tingting Fu[a], Yunan Ji[b], Kenju Kamei[c], Louis Putterman[d*]

[a]*Department of Economics, Nankai University, China*
[b]*Harvard University GSAS, United States*
[c]*Durham University Business School, United Kingdom*
[d]*Department of Economics, Brown University, United States*
*Corresponding author. Tel.: +1 401 863 3837. Email: Louis_Putterman@brown.edu

Abstract: Do opportunities to punish non-punishers help to stabilize cooperation? Or do opportunities to punish punishers harm cooperation and its benefits by deterring first order punishment and wasting resources? We compare treatments of a decision experiment without peer punishment and with one order of punishment to ones in which subjects can be punished for punishing or for failing to punish. Our treatments with higher-order punishment achieve as much improvement in cooperation as those with only one punishment stage. We see evidence of social norms in action, but no evidence of punishing failure to punish. These results suggest that higher-order punishment is neither critical to nor a major deterrent to cooperation.

Keywords: Punishment, cooperation, retaliation, higher order punishment

A lively discussion among evolutionary theorists addresses the problem of reconciling observed human cooperation with the drive to maximize reproductive fitness (Sober and Wilson, 1998, Boyd and Richerson, 2009, Axelrod and Hamilton, 1981). Many contributors assign a large role to social norms enforced by peer punishment (Fehr and Gächter, 2002, Henrich, 2004). But controversy exists over whether the punishability of punishment choices themselves—higher-order punishment—is helpful or harmful to cooperation. Whereas the theorists argue that individuals standing ready to punish those who omit to punish may be a key stabilizer of first-order punishment (Axelrod, 1986, Henrich and Boyd, 2001), some laboratory decision studies

have found that opportunities to engage in higher-order punishment are efficiency-reducing (Denant-Boemont *et al*., 2007, Nikiforakis, 2008).

We conduct an experiment comparing cooperation dilemmas without punishment opportunities or with only one order of punishment to ones permitting multiple orders of punishment under varying information conditions. In our main treatments, each of 240 subjects is grouped with three others in sessions of 20 participants. Each makes decisions on allocating funds between a private account and a group account in a standard voluntary contribution design of 15 periods in fixed groups. As with past experiments, selfish rational actors with common knowledge of type are predicted to put all tokens in their private accounts, but in line with past results subjects in the Baseline (no punishment) condition initially put about half of tokens in the group account, their average contribution then declining with repetition (Ledyard, 1995, Zelmer, 2003). In treatment Punish 1, we add now-standard opportunities to punish fellow group members after being shown their contributions, in a second stage of each period. Punishing, because costly to the punisher, is not predicted of selfish rational actors, but like past studies the treatment generates much punishing, mostly directed at lower contributors and mostly from higher ones, and contributions are significantly higher than Baseline and show a rising rather than declining trend until the final period (Fehr and Gächter, 2000, Gächter *et al*., 2008).

In our third treatment, Punish 2, each period has a third stage in which subjects learn the amount and originator of any punishment they received in the second one and can spend money to counter-punish. In their similar treatments, Denant-Boemont *et al.* (2007) and Nikiforakis (2008) found substantial counter-punishment, decline in first-order punishing, and more decline

of contributions. Consistent with the first finding, our subjects counter-punish 29% of punishing events. Nevertheless, average contributions are also significantly higher than Baseline in Punish 2, and there is no significant difference in contributions or earnings between Punish 1 and Punish 2. 83.3% of first-order punishing in Punish 1 and 95% in Punish 2 go to below-average contributors (two-tailed group-level Mann-Whitney test, z=-1.060, p=0.2892). As we show in the Appendix, in both treatments, below-average contributors increase their contribution from one period to the next by a larger amount the more punishment they receive. The pattern of counter-punishing also suggests presence of implicit norms: in Punish 2, a unit of first-order (stage 2) punishment given by a lower to a higher contributor (antisocial punishment in the terminology of Herrmann *et al*. 2008) triggers an average of 0.56 units of counter-punishment (in stage 3), whereas a unit of punishment from a higher to a lower contributor (prosocial punishment) leads to 0.24 units of counter-punishment.

Punish 2 allows a punished individual *j* to punish back her punisher *i*, but information about punishing or lack thereof between other pairs of group members is not made available, and third-party enforcement, such as punishing those who fail to punish low contributors or those who punish high contributors, is ruled out. We explore these omissions by conducting Punish 2´, a treatment in which we show subjects information about all punishments, then allow them to punish in the period's second punishment stage without restriction, keeping fixed subject identifiers for all periods a group interacts. Subjects are also shown a reminder of the previous period's contributions and punishments and of average contribution and punishments of each group member and dyad in periods before that, easing demands on memory. Figure 1 shows that the pattern of contributions over time in Punish 2´ is similar to the patterns in Punish 2 and

Punish 1. Figure 2 compares the average contribution and earnings across the four treatments: contributions are significantly higher in Punish 2´ than in Baseline, but there is no statistically significant difference in contribution between Punish 2´ and either Punish 2 or Punish 1. Average earnings are higher in each treatment allowing punishment than in Baseline. Although these pairwise earnings differences are statistically significant only for Punish 1 ($p = 0.031$) and Punish 2´ ($p = 0.065$), there are no statistically significant differences in earnings between any two punishment treatments, meaning presence of an additional punishment stage in Punish 2 and Punish 2´ does not significantly lower earnings relative to Punish 1. These findings are corroborated using random effects tobit and ordered probit regressions, as shown in our Appendix.

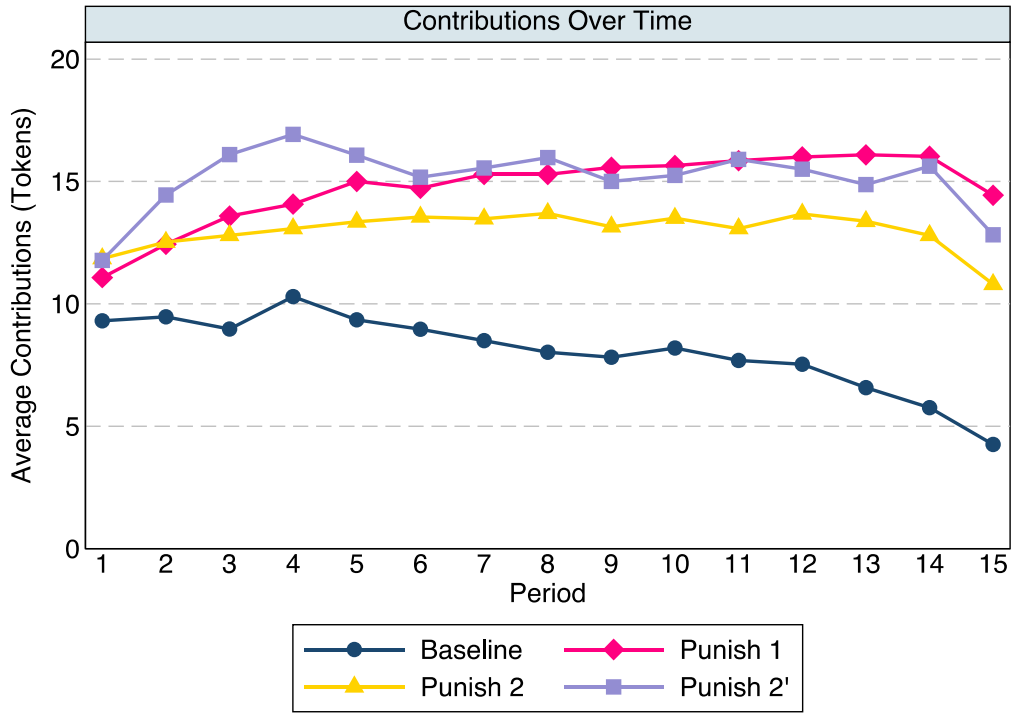

Fig. 1. Average amount contributed, out of 20 tokens, by period and treatment. See text for treatment descriptions.

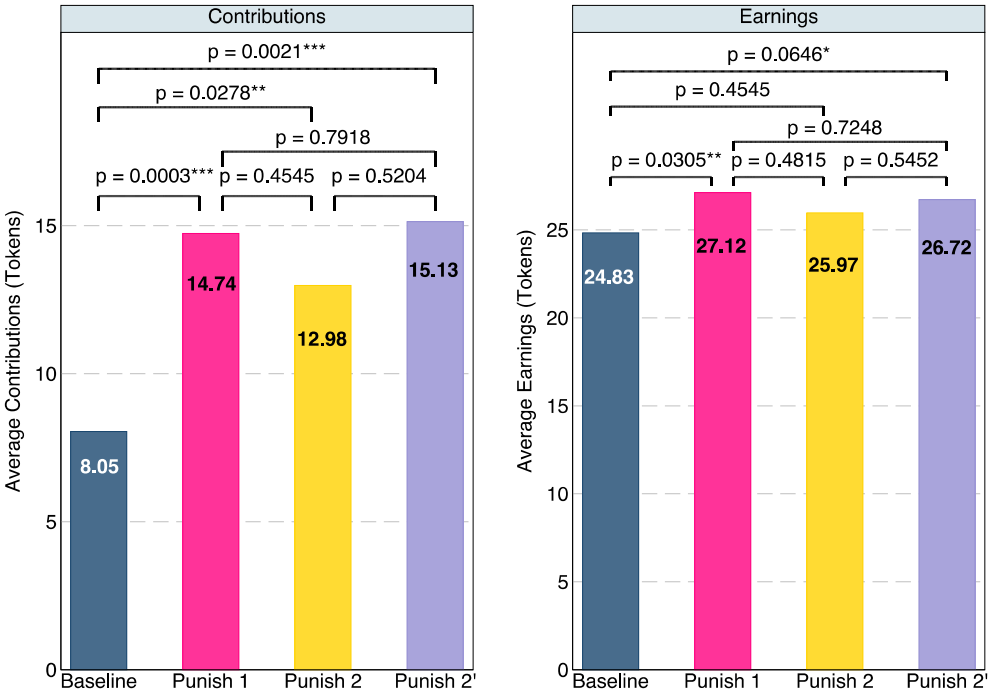

Fig. 2. Average contribution (in the left panel) and average earnings (in the right panel), by treatment. *p*-values from two-tailed Mann-Whitney Tests of group-level average contribution across all 15 periods are shown at the top of each panel. N = 20 groups (80 subjects) in Baseline and Punish 1, 10 groups (40 subjects) in Punish 2 and Punish 2´. * p < 0.1 ** p < 0.05 *** p < 0.01

As in Punish 2, Punish 2´ manifests costly punishment and costly counter-punishment, the latter occurring on average in 41% of instances of first-order punishment, not statistically significantly different from the 29% in Punish 2 (two-tailed group-level Mann-Whitney test, z = -0.090, p = 0.9283). The pattern of counter-punishment is similar, with 0.85 units of counter-punishment per unit when the punishment is antisocial, versus 0.42 units when it is prosocial. 56% (62%) of antisocial punishment events draw counter-punishment in Punish 2´ (Punish 2), whereas only 24% (28%) of prosocial punishment events are counter-punished. The differences in frequency

of counter-punishment when punishment is prosocial versus antisocial are statistically significant in the linear regression models of Table 1 and similar ordered logit regression models (see Appendix Table A9). A pooled linear regression model (see Table A10) shows that frequency of counter-punishment of each type does not significantly differ between Punish 2 and Punish 2´.

	Dependent variable: Counter-punishment event dummy					
	(i)	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**
	Punish 2	**Punish 2´**	**Punish 2**	**Punish 2´**	**Punish 2**	**Punish 2´**
Antisocial punishment (dummy)	0.235***	0.358***				
	(0.086)	(0.061)				
Max $\{(C_{jt} - C_{it}), 0\}\times$ antisocial punishment dummy			0.050***	0.016***		
			(0.007)	(0.004)		
Max $\{(C_{jt} - C_{it}), 0\}\times$ (P_{ijt} if $C_{jt} \geq C_{it}$, else 0)					0.044***	0.003**
					(0.008)	(0.001)
Constant	0.290***	0.203***	0.287***	0.308***	0.280***	0.339***
	(0.050)	(0.076)	(0.045)	(0.090)	(0.041)	(0.094)
Observations	118	75	118	75	118	75
Number of subjects	27	25	27	25	27	25
R-squared	0.0628	0.1118	0.0699	0.0121	0.0852	0.0022
Wald Chi-squared	7.49	34.56	54.84	16.74	27.28	5.74
Prob > Chi-squared	0.0062	0.0000	0.0000	0.0000	0.0000	0.0166

Table 1. Linear probability regression model of probability of counter-punishment as a function of type of first-order punishment. Regressions include all period-subject pair observations in which *i* punished *j* in stage 2 of the period. The dependent variable is 1 if *j* punished *i* back in stage 3, else 0. In column (i) and (ii), antisocial punishment takes value 1 if *j*'s contribution in

the period was greater than or equal to *i*'s, else 0. Regressions include individual random effects and group-level clustering of errors. The result shows that counter-punishment occurs with probability of about 0.3 or 0.2 when punishment is prosocial, versus probabilities of about 0.53 or 0.56 when punishment is antisocial, with difference of counter-punishment incidence being statistically significant at the 1% level in each treatment. In column (iii) and (iv), the independent variable is an interaction term of the antisocial punishment dummy variable and the positive deviation of *j*'s contribution from *i*'s, arguably a measure of "how antisocial" the punishment event was. In column (v) and (vi), the independent variable is the product of the previous specification's deviation term and the amount of punishment *i gave to j*, hence modifying the column (iii) and (iv) variable to also account for the amount of punishment *j* is reacting to. When we add to the independent variables the total 1st order punishment received by *j*, the coefficients and significance levels of the main explanatory variables remain the same, while the coefficient on the additional variable is not significant. Standard errors in parentheses. ** $p < 0.05$ *** $p < 0.01$

Regression analysis (see Table A4 and other Appendix tables) makes clear that the more a subject punishes in the first punishment stage, the more he or she is punished in the second. This holds both when first order punishment is prosocial and when it is antisocial. Indicator variables for giving no punishment or for failing to punish a low contributor when the opportunity arises are negatively associated with punishment received in the period's final stage, significantly so in several specifications (see Appendix Tables A5 – A7). As an additional check, we conduct a treatment resembling Denant-Boemont *et al*.'s No Revenge treatment, wherein subjects are shown only punishing not directed at themselves. Here, too, regressions indicate that giving no

first order punishment is if anything negatively associated with second order punishment received (see Appendix discussion of the Punish 2k treatment). Together, these results imply that the least higher-order punishment goes to non-punishers, hence higher-order punishment provides no net inducement to punish.

We conclude that whereas some experimental studies of cooperation have questioned whether the apparent benefits of peer punishment for cooperation can withstand the availability of higher-order punishment, our study of such punishment in a new laboratory experiment finds that the positive impact of first order punishment opportunities on cooperation is not an artifact of punishers being protected from higher-order punishment. Higher-order punishment does not significantly affect the cooperation-inducing effects of first-order punishment, among our subjects.[1] By the same token, whereas some evolutionary theorists have hypothesized that punishment of those who free-ride on punishment might explain how a tendency to punish could have evolved and stabilized, higher-order punishment shows no sign of encouraging first-order punishment in our experiment.

[1] Conceivably, a source of difference between our results and those of Denant-Boemont *et al*. (2007) and Nikiforakis (2008) is that our subjects are students at universities in Tianjin, China, rather than Europe. In the Appendix, we investigate this issue in detail with data from Herrmann *et al*. (2008) to find that the Tianjin subjects' behaviors are quite like those of European and U.S. counterparts in most respects. We also discuss the differences of our results from those of Kamei and Putterman (2015), a related U.S. experiment that attains qualitatively similar results to those authors in a treatment resembling Punish 2, but differences with both those authors and our present results in a treatment resembling Punish 2´.

References

Axelrod R, Hamilton WD (1981) The evolution of cooperation. *Science.* 211(27):1390-1396.

Axelrod R (1986) An evolutionary approach to norms. *Am. Pol. Sc. Rev.* 80(04):1095-1111.

Boyd R, Richerson PJ (2009) Culture and the evolution of human cooperation. *Phil. Trans. R. Soc. B.* 364(1533):3281-3288.

Denant-Boemont L, Masclet D, Noussair CN (2007) Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Econ. Theor.* 33(1):145-167.

Fehr E, Gächter S (2000) Cooperation and Punishment in Public Goods Experiments. *Am. Econ. Rev.* 90(4):980-994.

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

Gächter S, Renner E, Sefton M (2008) The Long-run Benefits of Punishment. *Science.* 322(5907):1510-1510.

Henrich J, Boyd R (2001) Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *J. Theor. Biol.* 208(1):79-89.

Henrich J (2004) Cultural group selection, coevolutionary processes and large-scale cooperation. *J. .Econ. Behav. Organ.* 53(1):3-35.

Herrmann B, Thöni C, Gächter S (2008) Antisocial punishment across societies. *Science.* 319(5868):1362-1367.

Kamei K, Putterman L (2015) In broad daylight: fuller information and higher-order punishment opportunities can promote cooperation. *J. Econ. Behav. Organ.* 120:145-159.

Ledyard J (1995), “Public Goods: A Survey of Experimental Research” in *Handbook of Experimental Economics*, J. Kagel, A. Roth, Eds. (Princeton University Press, Princeton, NJ), pp 111-194.

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

Sober E, Wilson DS (1998) *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Harvard Univ. Press, Cambridge, MA).

Zelmer J (2003) Linear public goods experiments: a meta-analysis. *Exp. Econ.* 6(3):299-310.

Acknowledgments: The authors thank He Jingtong of Nankai University for making it possible for the experiment to be conducted and for helping with the experimental process.