Kobe University Repository : Kernel

PDF issue: 2025-12-04

Does dishonesty really invite third-party punishment? Results of a more stringent test

Konishi, Naoki
Ohtsubo, Yohsuke

(Citation)
Biology Letters,11(5):20150172-20150172

(Issue Date)
2015-05-01

(Resource Type)
journal article

(Version)
Accepted Manuscript

(Rights)
©2015 The Author(s)

(URL)
https://hdl.handle.net/20.500.14094/90003504

Does Dishonesty Really Invite Third-Party Punishment?

Results of a More Stringent Test

Naoki Konishi Yohsuke Ohtsubo*

(Kobe University)

Department of Psychology, Graduate School of Humanities, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe, 657-8501, Japan

**author for correspondence (yohtsubo@lit.kobe-u.ac.jp).*

***Accepted for Publication in* Biology Letters**

Abstract

Many experiments have demonstrated that people are willing to incur cost to punish norm violators even when they are not directly harmed by the violation. Such altruistic third-party punishment is often considered an evolutionary underpinning of large-scale human cooperation. However, some scholars argue that previously demonstrated altruistic third-party punishment against *fairness*-norm violations may be an experimental artefact. For example, envy-driven retaliatory behaviour (i.e. spite) towards better-off unfair game players may be misidentified as altruistic punishment. Indeed, a recent experiment demonstrated that participants ceased to inflict third-party punishment against an unfair player once a series of key methodological problems were systematically controlled for. Noticing that a previous finding regarding apparently altruistic third-party punishment against *honesty*-norm violations may have been subject to methodological issues, we used a different and what we consider to be a more sound design to evaluate these findings. Third-party punishment against dishonest players withstood this more stringent test.

Keywords: third-party punishment; norms of honesty; experimental artefacts

1. Introduction

Large-scale cooperation among unrelated individuals characterises human sociality [1-3]. One explanation for the evolution of human cooperation is strong reciprocity, which assumes that people have inclinations for 'unconditional cooperation' and 'altruistic punishment against norm violators' [1]. Accumulated experimental evidence indicates that people incur some cost to punish norm violators even when they themselves or their relatives are not harmed by the norm violation [4-9]. Such altruistic third-party punishment has been observed among young children [6] and in many preliterate societies [7, 8], let alone in conventional adult samples in modern societies. Nonetheless, some scholars recently cast serious doubt on the existence of altruistic punishment [10, 11].

In a typical third-party punishment experiment [5], the participant acting as a third-party (Player C) observes the dictator game played by two players (Players A and B). Player A (the dictator), endowed with 100 points, is asked whether to give some of his/her endowment to Player B, who has no endowment. The participant (Player C), endowed with 50 points, is informed that he/she can reduce Player A's points by spending his/her own endowment. Specifically, if Player C spends x points, $3x$ points will be subtracted from Player A's points. Player C decides whether (and how much) to punish Player A assuming that Player A makes each of the six possible choices (i.e. giving from 0 to 50 points in increments of 10 points). Thereafter, Player C's points are determined by matching his/her decisions to Player A's actual choice. This so-called *strategy method* is widely used because it allows researchers to assess how Player C would behave in response to every possible situation ranging from fair to very unfair. Participants typically indicate willingness to incur some cost to punish Player A's unfair behaviours.

Pedersen and colleagues [11] maintain that many purported demonstrations of altruistic third-party punishment may be the result of experimental artefacts, as most third-party punishment experiments are associated with at least one of five common methodological problems. The first problem is limited behavioural choices—participants must choose whether to punish the unfair

player or do nothing at all. Therefore, if participants want to do *anything*, this will result in positive evidence for third-party punishment. The second problem is an audience effect [12]. Even implicitly assuming that someone (e.g. Player B) observes their behaviour, participants might inflict third-party punishment out of reputational concerns (e.g. to signal a fair image). The third and fourth problems are associated with the strategy method. Being exposed to both fair and unfair outcomes, participants might infer and comply with the research hypothesis. In addition, in order to make decisions under the strategy method, participants have to imagine how they would feel when Player A makes unfair decisions. However, people are bad at affective forecasting [13]. Therefore, participants might erroneously consider that they would punish the unfair player by overestimating their anticipated anger. The fifth problem is about the proximate emotion for third-party punishment. Researchers usually assume that anger causes third-party punishment [5, 14]. However, envy can cause Player C to behave in a spiteful manner towards an unfair Player A who earns more money than Player C. Thus, spite might be misidentified as punishment. Pedersen et al. conducted a series of third-party punishment experiments systematically controlling for these methodological problems, and showed that Player C, as compared to Player B (i.e. the victim of unfairness), rarely punished unfair players. Furthermore, the rare instances of ostensible third-party punishment were best accounted for by envy.

There is a study [9] that observed third-party punishment against *dishonesty*, instead of unfairness, which precluded three of the five methodological problems. In the study, a dishonest player lied to a second player by exaggerating his/her generosity, but then behaved in a fair manner (figure 1). When third-parties decided whether to punish the deceptive player, all players in the game possessed the same amount of money, thereby negating potential envy (problem 5). In addition, this study did not use the strategy method (problems 3 and 4). Nonetheless, 53% of participants in the study punished the deceptive player. As encouraging as this result is to the third-party punishment literature, it is still possible that the result was an artefact due to problems 1 and 2 (i.e. limited choice

and reputational concerns). Therefore, in the present study, we tested whether third-party punishment against dishonesty would be replicated after removing all five methodological problems. In particular, the present study added the reward option and the anonymity instructions to the previous study [9].

2. Method

(a) Participants and design

Participants were 83 (45 males and 38 females) undergraduates at a large university in Japan. We decided to discard 17 participants from the data analyses (33 participants in each condition were retained). Two participants failed to understand the payoff structure, and another participant personally knew the experimenter. Additional 14 participants were discarded because of their responses during the debriefing session. In specific, three participants spontaneously said that they suspected the absence of other players at the beginning of the debriefing session. In addition, towards the end of the debriefing session, the experimenter directly asked participants if they had even a little doubt about the presence of the other players. Eleven participants (six and five in the dishonesty and honesty conditions, respectively) gave a solid ‘yes’ to this direct inquiry, and were excluded. However, thirteen other participants who gave reserved forms of affirmative responses (e.g. ‘just a little bit’) were retained.

(b) Transactions between other players

Participants witnessed a modified version of the trust game, which involved a trustor and the trustee (figure 1). The two players first received an initial endowment of 500 Japanese yen (JPY). If the trustor decided to transfer his/her endowment to the trustee, it was tripled. The trustee then decided how to allocate 2,000 JPY. Unlike the standard trust game, this modified game allowed the trustee to send a pre-play message to the trustor. The message read either ‘I will give you 1,000 JPY’ in the honesty condition or ‘I will take 700 JPY and give you 1300 JPY’ in the dishonesty condition. After receiving the message, the trustor transferred his/her endowment. The trustee equally split 2,000 JPY in both conditions. Therefore, in the dishonesty condition, the trustee violated the honesty norm,

while complying with the fairness norm.

(c) Punishment

After observing either an honest or dishonest transaction, participants rated their current feelings towards the two players (see the electronic supplementary materials for details). Participants, who were endowed with 1,000 JPY, were then informed that they could increase or decrease the trustee's payoff by 2*c* JPY (where *c* stands for the cost that participants incur). Before making this decision, in order to minimise the audience effect [12], participants were ensured that both the trustor and the experimenter would be kept ignorant of their decision. In particular, participants were informed that another experimenter who would never see them would check their decision and prepare their experimental rewards, so that the experimenter whom they were meeting would be kept ignorant of their decision. Participants indicated whether or not they were willing to punish or reward the trustee (and if so, how much).

(d) Hypotheses

It was hypothesised that participants punish the dishonest player even after removing all five methodological problems. In particular, the following three propositions were tested: (1) there would be more punishers in the dishonesty condition than the honesty condition; (2) the cost participants incur would significantly differ from 0 (i.e. no punishment); (3) there would be more punishers than reawrders in the dishonesty condition.

3. Results

There were 9 punishers, 2 rewarders, and 22 unresponsive onlookers in the dishonesty condition, and 2 punishers, 7 rewarders, and 24 unresponsive onlookers in the honesty condition (figure 2). Removing two additional methodological problems (i.e. problems 1 and 2) from the original research [9], which already precluded the other three problems, reduced the punishment rate from 53% to 27% (= 9/33) in the dishonesty condition. Nonetheless, there were significantly more punishers in the dishonesty condition than in the honesty condition (6%): $p = .044$ by Fisher's exact

test.

We then tested whether the cost participants incurred to punish dishonesty exceeded 0. To compute the average cost, unresponsive onlookers were assigned 0, and rewarders' costs were assigned negative values (e.g. −100 for a participant who paid 100 JPY to reward the trustor). The average cost to punish dishonesty, 25.76 JPY (*sd* = 67.18), significantly deviated from 0 in the punitive direction, $t_{32} = 2.20$, $p = .035$. The comparable *t*-test indicated that the mean cost, −7.58 (39.77), was not significantly different from 0 in the honesty condition, $t_{32} = -1.09$, *ns*.

In the dishonesty condition, participants were more inclined to punish, rather than reward, the trustee (9 punishers vs. 2 rewarders: $p = .033$ by a binomial test with the assumption that participants were not biased to punish the dishonest trustee).

4. Discussion

The results indicated that removing all five methodological problems [11] did not completely eradicate third-party punishment against dishonesty. Although the methodological changes (i.e. the reward option and strengthened anonymity) substantially reduced the frequency of punishers, compared with the original study [9], the first analysis showed that there were still more punishers in the dishonesty condition than in the honesty condition. It is noteworthy that the reward option was a viable choice in the present study. If participants had been only concerned about the fairness norm, rewarding the fair but dishonest player, who divided 2,000 JPY equally, could have been a reasonable thing to do. Nevertheless, the second and third analyses showed that participants were more likely to punish, rather than reward, the fair but dishonest player.

Violations of the honesty norm might more reliably induce third-party punishment than violations of the fairness norm. For one thing, there are many children's moral stories that teach the virtue of honesty (e.g. The Boy Who Cried Wolf) [15]. In addition, punishment might be especially important for the honesty norm because of its association with linguistic communication, which is an instance of a metabolically cheap but honest signalling system. It has been shown that punishment

against dishonesty is crucial to keep such cheap signalling systems being evolutionarily stable [16].

In sum, third-party punishment against honesty norm violators cannot be completely dismissed as an experimental artefact. We agree that there is scant evidence for third-party punishment in general in real-life settings [10]. Nevertheless, some norms might have been more crucial than other norms for human groups to survive. Systematic considerations of the adaptive values of different norms seem to be needed to fully understand third-party punishment.

Ethics statement. This study was approved by the institutional review board at the corresponding author's institute.

Data accessibility statement. The data used in the reported analyses have been uploaded as the electronic supplementary material.

Competing interests statement. We have no competing interests.

Authors' contributions statement. N. Konishi conducted the experiment, analysed the data, wrote the first draft of the manuscript, and approved the final version of the manuscript. Y. Ohtsubo designed the experiment, analysed the data, revised the first draft, and approved the final version.

Acknowledgements. We are grateful to Daiki Inoue, Koji Kandori, Keisuke Matsugasaki, Haruna Okamoto, Adam Smith, Hiroki Tanaka, Kanako Tanaka, Kodai Tomita, Honoka Wada, Noriko Wada, Ayano Yagi, Chiaki Yamaguchi, and Ye-Yun Yu for their assistance.

Funding statement. We have received no specific grant for this study.

References

1. Gintis H. 2000 Strong reciprocity and human sociality. *J. Theor. Biol.* **206**, 169-179. (doi:10.1006/jtbi.2000.2111)

2. Fehr E, Fischbacher U. 2004 Social norms and human cooperation. *Trends Cogn. Sci.* **8**, 185-190.

(doi: 0.1016/j.tics.2004.02.007)

3. Boyd R, Gintis H, Bowles S, Richerson PJ. 2003 The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. USA* **100**, 3531-3535. (doi:10.1073/pnas.0630443100)

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

5. Fehr E, Fischbacher U. 2004 Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63-87. (doi:10.1016/S1090-5138(04)00005-4)

6. McAuliffe K, Jordan JJ, Warneken F. 2015 Costly third-party punishment in young children. *Cognition* **134**, 1-10. (doi:10.1016/j.cognition.2014.08.013)

7. Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, Bolyanatz A, et al. 2006 Costly punishment across human societies. *Science* **312**, 1767-1770. (dio:10.1126/science.1127333)

8. Marlowe FW, Berbesque JC, Barr A, Barrett C, Bolyanatz A, Cardenas JC, Ensminger J, Gurven M, Gwako E, Henrich J, et al. 2008 More ‘altruistic’ punishment in larger societies. *Proc. R. Soc. B* **275**, 587-590. (doi:10.1098/rspb.2007.1517)

9. Ohtsubo Y, Masuda F, Watanabe E, Masuchi A. 2010 Dishonesty invites costly third-party punishment. *Evol. Hum. Behav.* **31** 259-264. (doi:10.1016/j.evolhumbehav.2009.12.007)

10. Guala F. 2012 Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* **35**, 1-15. (doi:10.1017/S0140525X11000069)

11. Pedersen EJ, Kurzban R, McCullough ME. 2013 Do humans really punish altruistically? A closer look. *Proc. R. Soc. B* **280**, 1471-2954. (doi:10.1098/rspb.2012.2723)

12. Kurzban R, DeScioli P, O’Brien E. 2007 Audience effects on moralistic punishment. *Evol. Hum. Behav.* **28**, 75-84. (doi:10.1016/j.evolhumbehav.2006.06.001)

13. Wilson TD, Gilbert DT. 2003 Affective Forecasting. *Adv. Exp. Soc. Psychol.* **35**, 345-411. (doi:10.1016/S0065-2601(03)01006-2)

14. Seip EC, Van Dijk WW, Rotteveel M. 2014 Anger motivates costly punishment of unfair

behavior. *Motiv. Emot.* **38**, 578-588. (doi:10.1007/s11031-014-9395-4)

15. Lee K, Talwar V, McCarthy A, Ross I, Evans A, Arruda C. 2014 Can classic moral stories promote honesty in children? *Psychol. Sci*. **25**, 1630-1636. (doi:10.1177/0956797614536401)
16. Lachmann M, Szamado S, Bergstrom CT. 2001 Cost and conflict in animal signals and human language. *Proc. Natl. Acad. Sci. USA* **98**, 13189-13194. (doi:10.1073ypnas.231216498)

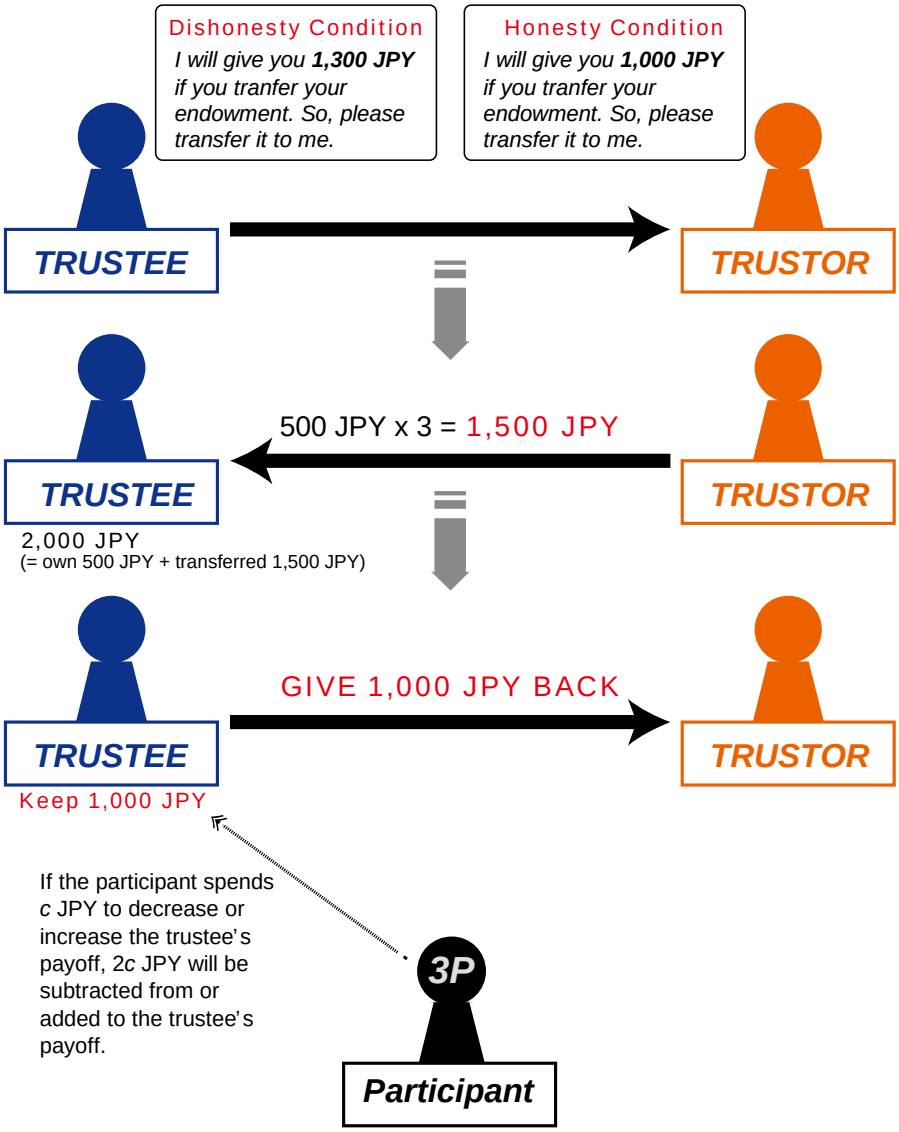

Figure 1. Schematic description of the experimental procedures. Participants (bottom of the figure) witnessed the transactions between the trustor and trustee, and decided whether to increase/decrease the trustee's payoff or do nothing.

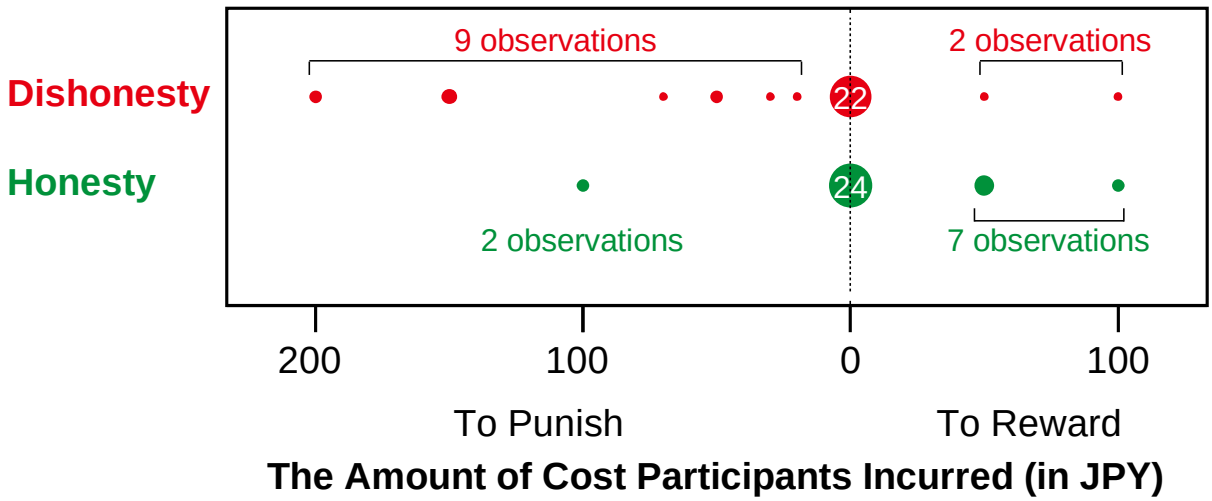

Figure 2. Bubble plot of the distribution of the cost that participants were willing to incur to increase/decrease the trustee's payoff (the size of circles represents the frequency of each data point). The left-side corresponds with punishment, and the right-side corresponds with reward.

Electronic Supplementary Materials

Does Dishonesty Really Invite Third Party Punishment?

Results of a More Stringent Test

Naoki Konishi Yohsuke Ohtsubo

(Kobe University)

METHOD

Emotions towards the trustee

After observing either honest or dishonest transactions, participants rated how they were currently feeling towards each of the two players. The emotion questionnaire consisted of five anger items (i.e. angry, indignant, perturbed, mad, dissatisfied), two envy items (i.e. envious, jealous), and five empathy items (i.e. empathetic, warm, tender, sympathetic, concerned), which were adapted from previous studies with modifications [1, 2]. Participants rated their feelings on a 6-point scale (0 = 'not at all' to 5 = 'very much').

Cronbach's α coefficient for anger was .95 and .81 for the trustee and the trustor, respectively. The correlation between the two envy items was significant but small (r_{64} = .29 and .26 for the trustee and the trustor, respectively). Therefore, we decided to use the responses to the envy-item the envy score, instead of aggregating the responses to the envy- and jealousy-items. For the empathy items, two of the five items were not highly correlated with the other items. Cronbach's α coefficient based on the remaining three items (empathetic, warm, tender) was .82 and .75 for the trustee and trustor, respectively (cf. the five-item α was .69 and .33 for the trustee and the trustor, respectively).

Other details of the procedure

Participants were informed that they were able to increase or decrease the trustee's reward by spending some of their own reward for the participation. At the beginning of the experiment,

participants were shown a tick ‘filler’ questionnaire, and explained that they would be paid 1,000 Japanese yen (JPY) for completing the questionnaire after a game task. This brief procedure was included in order to minimise the possibility of demand characteristics. If participants believed that the only task they would engage in was to decide whether to make some alteration to the trustee’s reward, they might feel obliged to do so.

Participants were explained that they must pay the cost of *c* JPY in order to increase or decrease the trustee’s reward by 2*c* JPY. Before participants made their decision, in order to minimise a potential audience effect, participants were explicitly informed that the trustor would not be told whether they increased/decreased the trustee’s reward (see also the main text). In addition, participants were explained their decision would not at all influence the trustor’s reward.

Participants then decided whether or not to make any change to the trustee’s reward (‘yes’ or ‘no’). When the answer to this first question was ‘yes’, they were asked whether to increase or decrease the trustee’s reward. Finally, they indicated the cost (*c*) that they would incur to make the indicated alteration. After their decision, participants were asked to write down the final rewards of the three players. This last inquiry served as the check of participants’ correct understanding of the payoff structures of this game. Two participants in the honesty condition made mistakes on this comprehension check, and were omitted from the subsequent data analyses.

RESULTS

Sex Differences in Third-party punishment

There were five (out of 15) male punishers and four (out of 18) female punishers in the dishonesty condition. Males and females did not differ in their punitive tendency, $p = .697$, by Fisher’s exact test.

Emotions towards the trustee

As shown in table S1, participants’ anger and envy scores were not particularly high across

the conditions and the targets (cf. the range of these scores was 0 to 5). Nonetheless, when these scores were submitted to 2 (condition) × 2 (target) ANOVA, the condition × target interaction effect was significant, $F_{1, 64} = 22.86$, $p < .001$ for anger and $F_{1, 64} = 6.99$, $p = .010$ for envy, along with the significant main effects of condition ($F_{1, 64} = 22.62$, $p < .001$) and target ($F_{1, 64} = 23.84$, $p < .001$) for anger and at least marginally significant main effects of condition ($F_{1, 64} = 3.61$, $p = .062$) and target ($F_{1, 64} = 8.98$, $p = .004$) for envy. Post hoc comparisons by Ryan's method indicated that participants were particularly angry at and envious of the dishonest trustee.

Participants' empathy scores were also submitted to the 2 (condition) × 2 (target) ANOVA. The interaction effect ($F_{1, 64} = 19.53$, $p < .001$) along with the two main effects ($F_{1, 64} = 33.13$, $p < .001$ for condition and $F_{1, 64} = 8.10$, $p = .006$ for target) were significant. Post hoc comparisons revealed that participants were especially less empathetic with the dishonest trustee. However, in this case, it was also shown that participants were less empathetic with the trustor in the dishonesty condition. It might be an instance of victim derogation (participants might have re-interpreted the trustor's trustful decision as being stupid or too naive) [3].

Table S1. Mean scores (*SD*) of the three emotions towards the trustee and the trustor as a function of the trustee's behaviour.

Emotion towards	Condition	Anger	Envy	Empathy
Trustee	Dishonesty (n = 33)	1.34 (1.30)	0.54 (0.97)	1.43 (1.06)
	Honesty (n = 33)	0.14 (0.28)	0.12 (0.33)	3.31 (1.11)
Trustor	Dishonesty (n = 33)	0.18 (0.47)	0.06 (0.24)	2.41 (1.08)
	Honesty (n = 33)	0.13 (0.32)	0.09 (0.29)	3.10 (0.98)

The above analyses indicate that the presence of dishonesty induced little (although significant) anger and envy. On the other hand, empathic concern with the trustee was substantially reduced by observing the trustee's dishonest behaviour. We then tested whether these emotional reactions were correlated with participants' punishment behaviour (a dichotomous variable: 0 = 'not punished' and 1 = 'punished'). As can be seen in Table S2, lowered empathy with the trustee and trustor was significantly correlated with punishment behaviour. If victim derogation was in fact the cause of lowered empathic concern with the trustor, the correlation between punishment and empathy towards the deceived trustor can be a spurious correlation mediated by the trustee's dishonest behaviour (i.e. the victimisation of the trustor). When the dichotomous punishment variable was regressed on empathy towards the trustee and the trustor, lowered empathy towards the trustee significantly predicted punishment ($\beta = -.41$, $p = .006$), but lowered empathy towards the trustor did not ($\beta = -.08$, *ns*). This result is consistent with recent evidence that lowered empathy towards perpetrators is a proximate cause of punishment [4, 5].

Table S2. Correlation coefficients between the emotion scores and punishment behaviour (the bold and red fonts indicate statistical significant at the .05-level).

		Anger		Envy		Empathy	
		trustee	trustor	trustee	trustor	trustee	trustor
Punishment		.20	.24	.07	.03	**−.45**	**−.30**
Anger	trustee		.15	**.64**	.01	**−.62**	−.07
	trustor			.23	**.47**	−.09	−.09
Envy	trustee				**.26**	**−.29**	−.05
	trustor					.11	.03
Empathy	turstee						**.54**

We then tested whether the empathy towards the trustee mediates the effect of condition (dishonesty vs. honesty) on third-party punishment. As shown in figures S1, the Sobel test indicated that empathy towards the trustee fully mediated the effect of the condition (z = 2.85, p = .002). Previous studies [4, 5] have shown that empathy is associated with third-party help of the victim as well. In future research, behavioural options must be expanded so that participants can freely choose whether to punish the transgressor or help the victim. In previous studies, empathy towards the victim (the trustor in the present study) caused participants to help the victim. However, the present study suggests the presence of victim derogation, which might undermine the positive effect of empathy on helping.

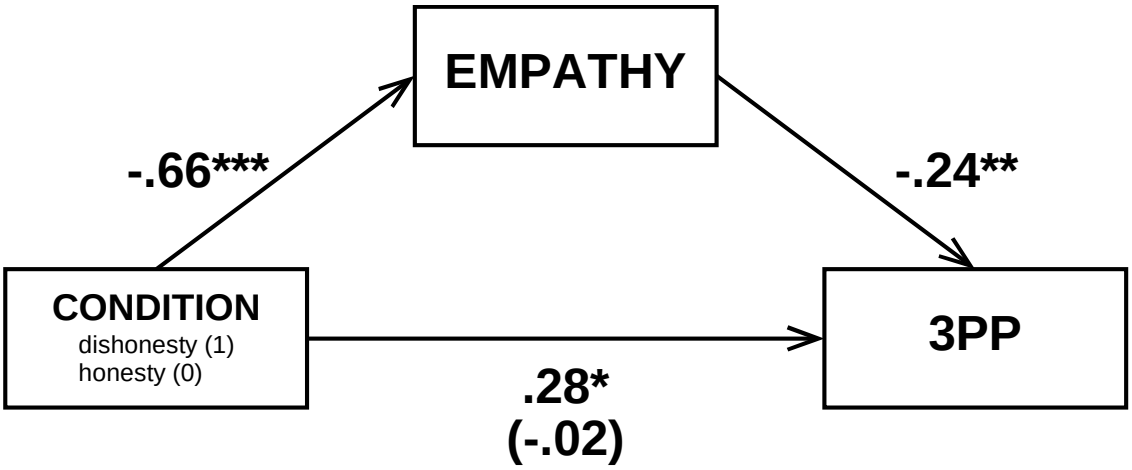

Figure S1. Results of mediation analysis. Empathy towards the trustee fully mediated the relationship between condition and third-party punishment.

References

1. Pedersen EJ, Kurzban R, McCullough ME. 2013 Do humans really punish altruistically? A closer look. *Proc. R. Soc. B* **280**, 1471-2954. (doi:10.1098/rspb.2012.2723)

2. Batson CD, Kennedy CL, Nord L-A, Stocks EL, Fleming DA, Marzette CM, Lishner DA, Hayes RE, Kolchinsky LM, Zerger T. 2007 Anger at unfairness: Is it moral outrage? *Eur. J. Soc. Psychol.* **37**, 1272-1285. (doi:10.1002/ejsp.434)

3. Lerner MJ, Simmons CH. 1966 The observer's reaction to the "innocent victim": Compassion or rejection? *J. Pers. Soc. Psychol.* **4**, 203-210. (doi:10.1037/h0023562)

4. Leliveld MC, Van Dijk E, Van Beest I. 2012 Punishing and compensating others at your own expense: The role of empathic concern on reactions to distributive injustice. *Eur. J. Soc. Psychol.* **42**, 135-140. (doi:10.1002/ejsp.872)

5. Hu Y, Strang S, Weber B. 2015 Helping or punishing strangers: neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Front. Behav. Neurosci.* **9**:24. (doi: 10.3389/fnbeh.2015.00024)