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

Naoki Konishi and Yohsuke Ohtsubo

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

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.

1. Introduction

Large-scale cooperation among unrelated individuals characterizes human society [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 et al. [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. 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 with 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 that 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

![Figure 1. Schematic of the experimental procedures. Participants in the third-party (3P) role witnessed the transactions between the trustor and trustee, and decided whether to increase/decrease the trustee’s payoff or do nothing. JPY, Japanese yen; c, the cost that participants incur. (Online version in colour.)](http://rsbl.royalsocietypublishing.org/)
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 this study, we tested whether third-party punishment against dishonesty would be replicated after removing all five methodological problems. In particular, this study added the reward option and the anonymity instructions to the previous study [9].

2. Method

(a) Participants and design

Participants were 83 (45 male and 38 female) 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. An additional 14 participants were discarded because of their responses during the debriefing session. Specifically, 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 2000 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 1000 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 2000 JPY in both conditions. Therefore, in the honesty condition, which already precluded the other three problems, reduced the punishment rate from 53 to 27% (i.e. 9/33) in the dishonesty condition. Nonetheless, there were significantly more punishers in the dishonesty condition than in the honesty condition (6%): $p = 0.044$ 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 trustee). The average cost to punish dishonesty, 25.76 JPY (s.d. = 67.18), significantly deviated from 0 in the punitive direction, $t_{32} = 2.20$, $p = 0.035$. The comparable $t$-test indicated that the mean cost, $−7.58 (39.77)$, was not significantly different from 0 in the honesty condition, $t_{32} = −1.09$, n.s.

In the dishonesty condition, participants were more inclined to punish, rather than reward, the trustee (nine punishers versus two rewarders: $p = 0.033$ by a binomial test with the assumption that participants were not biased to punish the dishonest trustee).

(c) Punishment

After observing either an honest or dishonest transaction, participants rated their current feelings towards the two players (see the electronic supplementary material for details). Participants, who were endowed with 1000 JPY, were then informed that they could increase or decrease the trustee’s payoff by 2c JPY (where $c$ stands for the cost that participants incur). Before making this decision, in order to minimize the audience effect [12], participants were assured 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 hypothesized that participants punish the dishonest player even after removing all five methodological problems. In particular, the following three propositions were tested: (i) there would be more punishers in the dishonesty condition than the honesty condition; (ii) the cost participants incur would significantly differ from 0 (i.e. no punishment); (iii) there would be more punishers than rewarders in the dishonesty condition.

3. Results

There were nine punishers, two rewarders and 22 unresponsive onlookers in the dishonesty condition; and two punishers, seven 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% (i.e. 9/33) in the dishonesty condition. Nonetheless, there were significantly more punishers in the dishonesty condition than in the honesty condition (6%): $p = 0.044$ 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 trustee). The average cost to punish dishonesty, 25.76 JPY (s.d. = 67.18), significantly deviated from 0 in the punitive direction, $t_{32} = 2.20$, $p = 0.035$. The comparable $t$-test indicated that the mean cost, $−7.58 (39.77)$, was not significantly different from 0 in the honesty condition, $t_{32} = −1.09$, n.s.

In the dishonesty condition, participants were more inclined to punish, rather than reward, the trustee (nine punishers versus two rewarders: $p = 0.033$ 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 this study. If participants had been only concerned about the fairness norm, rewarding the fair but dishonest player, who divided 2000 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 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.

References