Third-party punishers are rewarded - but third-party helpers even more so

Nichola J Raihani [1*] & Redouan Bshary [2]

1. Department of Genetics, Evolution and Environment, University College London, London UK. WC1E 6BT.

2. Institut de Biologie, Université de Neuchâtel, Switzerland CH-2000.

*Author for correspondence: nicholaraihani@gmail.com

Data will be made available as an additional supplementary online material file.

Punishers can benefit from a tough reputation, where future partners cooperate because they fear repercussions. Alternatively, punishers might receive help from bystanders if their act is perceived as just and other-regarding. Third-party punishment of selfish individuals arguably fits these conditions but it is not known whether third-party punishers are rewarded for their investments. Here, we show that third-party punishers are indeed rewarded by uninvolved bystanders. Third-parties were presented with the outcome of a dictator game where the dictator was either selfish or fair and were allocated to one of three treatments where they could choose to do nothing or (i) punish the dictator, (ii) help the receiver or (iii) choose between punishment and helping, respectively. A fourth player ('bystander') then saw the third-party's decision and could choose to reward the third-party or not. Third-parties that punished selfish dictators were more likely to be rewarded by bystanders than third-parties who took no action in response to a selfish dictator. However, helpful third-parties were rewarded even more than third-party punishers. These results suggest that punishment could in principle evolve via indirect reciprocity but also provide insights into why individuals typically prefer to invest in positive actions.

Key words: punishment, indirect reciprocity, reputation, cooperation

Introduction

Indirect reciprocity has been proposed as a solution to the problem of why individuals pay to help others in social interactions. Indirect reciprocity occurs when one individual helps another and improves his image score (Nowak and Sigmund 1998a, b) or standing (Sugden 1986) as a result. As a consequence of his helpful actions, the first individual is then more likely to be helped by an

uninvolved third-party. Theoretical and empirical studies have demonstrated the success of indirect reciprocity for favouring helping behavior when defecting would otherwise be more profitable (Nowak and Sigmund 1998a, b; Wedekind and Milinski 2000; Leimar and Hammerstein 2001; Milinski et al. 2001; Seinen and Schram 2006). In contrast, far less attention has been paid to the possibility that punishment allows individuals to improve their reputation and to be more likely to receive help from bystanders as a result (Raihani and Bshary 2015). We address this possibility here.

Punishment occurs when an individual invests to reciprocally harm a cheating partner (Clutton-Brock and Parker 1995; Raihani et al. 2012) . In two-player games, this investment can be recouped if the victim behaves more cooperatively in future interactions with the punisher. However, under some circumstances, the benefits of punishment are less obvious. For example, in n-player games, the punisher bears the cost of punishment while the benefits of increased cooperation within the group are shared among all group members (Fehr and Gächter 2002). In third-party punishment, the punisher invests to punish a cheat even when he was not harmed by the cheat's behavior and does not necessarily expect to interact with the cheat in future (e.g. Fehr and Fischbacher 2004). Theoretical and empirical work has demonstrated that punishment can evolve via individual-level benefits if the punisher acquires a punitive reputation (Brandt et al. 2003; dos Santos et al. 2011; 2013; Hilbe and Traulsen 2012). In a recent study, dos Santos et al. (2013) used a two-player helping game with punishment to show that individuals that punished a partner for refusing to help were more likely to be helped by bystanders in subsequent rounds. When punishment was prevented, however, individuals with a punitive reputation did not receive more help than those without, indicating that punishers did not receive voluntary help in the absence of a punishment threat (dos Santos et al. 2013). Thus, according to this study at least, punishers benefit from acquiring a punitive reputation because bystanders fear repercussions and so behave more cooperatively with the punisher, not because punishment is deemed worthy of reward. Other

empirical studies have shown similar patterns: punishers in public goods games are not rewarded (Kiyonari and Barclay 2008) and punishers may be less likely than non-punishers to be chosen for future interactions (Horita 2010; Ozono and Watabe 2012).

Despite these empirical findings, it has recently been suggested that punishment might sometimes be approved of by bystanders, with punishers receiving voluntary rewards or being chosen as a partner based on their punitive behaviour (Raihani and Bshary 2015). Specifically, bystanders should be sensitive to the motives underpinning punishment and be more likely to reward (or choose) punishers only when punishment reliably signals that the punisher acted out of concern for the welfare of others, rather than themselves (Raihani and Bshary 2015). It was argued that punishment that was apparently motivated by concern for others' welfare is more likely to act as a reliable signal of the punisher's cooperative intent. Where punishment reliably signals the punisher's fairness preferences or cooperative tendency, punishers may be evaluated in a similar manner to helpful individuals and therefore receive voluntary rewards or be preferred as interaction partners (Raihani and Bshary 2015).

Punishment is most likely to act as a signal that the punisher is concerned for the welfare of others if the punisher was not the initial victim of the cheat and does not stand to benefit from any change in the cheat's behavior (Raihani and Bshary 2015): this is often the case with third-party punishment. Third-party punishment is typically studied in the context of a one-shot game, where the punisher is given the option to punish a cheat that harmed another individual (e.g. Kurzban et al. 2007; Nelissen 2008; Piazza and Bering 2008). Since the third-party is not the victim of the cheat and will not interact with the cheat in future, bystanders might infer that the punisher acted out of a concern for the welfare of the victim (but see Pedersen et al. 2013). By contrast, where the punisher was the victim of the cheat or will interact with the cheat again in future rounds, bystanders may infer that punishment was motivated by the punisher's negative emotions at experiencing loss or

inequity as a consequence of interacting with the cheat (e.g. Fehr and Gächter 2002; Raihani and McAuliffe 2012; Bone and Raihani, in press) or by the desire to change the cheat's behavior in future (e.g. Benard 2013). Punishment under these conditions is unlikely to signal the punisher's concern for others' welfare and bystanders might be more likely to infer that the punishment was driven by competitive rather than cooperative motives. Competitive behaviour is unlikely to be interpreted by bystanders as a behavior aimed at helping others and punishers are unlikely to be rewarded or preferred as partners under these circumstances (Raihani and Bshary 2015).

Some empirical evidence supports the idea that third-party punishers are viewed positively by others (e.g. Gordon et al. 2014): individuals are more likely to invest in third-party punishment if their investment will be advertised to others (Kurzban et al. 2007; Piazza and Bering 2008) and individuals that incur greater costs to deliver third-party punishment are preferred as interaction partners for subsequent trust games (Nelissen 2008). However, although third-party punishment indicates that punishers are trustworthy (Barclay 2006) and therefore preferable as partners for interactions requiring trust (Nelissen 2008), this does not necessarily imply that punishers will be approved of or receive voluntary rewards from others. For example, a recent study used vignettes to demonstrate that punishers are preferred for interactions when they can provide some benefit to the partner (e.g. when the punisher will be the dictator in a dictator game) but are less preferred when they will be in the beneficiary role (Horita 2010). Here, we ask whether third-party punishers receive voluntary rewards from bystanders. If third-party punishment improves an individual's reputation, we predict that bystanders will reward punishers more than non-punishers.

There is an increasing awareness that while various partner control mechanisms may promote cooperation in a world initially dominated by defectors, one should also ask which mechanism may be favoured over others (Bshary and Bronstein 2011). Theoreticians working on the reputation effects of punishment have addressed this issue, concluding that punishment of defectors typically

dominates rewarding contributors in reputation games (Brandt et al. 2003; dos Santos et al. 2011; Hilbe and Traulsen 2012; Sigmund et al. 2001; Hilbe and Sigmund 2010; Roos et al. 2014 but see Ohtsuki et al. 2009). However, the models are concerned with game structures in which punishment causes a reputation of fear, inducing partners to cooperate in order to avoid being punished themselves. In a pure indirect reciprocity scenario as tested here, help should be given voluntarily, without the threat of being punished otherwise. In a second step we therefore ask to what extent the reputation gained by third-party punishment compares to reputation gained by helping the victim, both in an independent comparison (third-parties only have one response option) and when third-parties can choose between punishing egoists and helping victims. Empirical studies suggest that third-party punishment may be more likely to serve as a reliable signal of concern for others if it is the only signalling option available (Raihani and Bshary 2015). Individuals may desist from punishing or hide punitive actions from bystanders when they can signal concern for others via positive actions such as cooperating or compensating victims (e.g. Charness et al. 2008; Chavez and Bicchieri 2013; Rockenbach and Milinski 2011). This is because, while the costs for the actor may be the same, punishment by definition destroys value and thus typically lowers population productivity (Ohtsuki et al. 2009) while helping increases value. Thus, we predict that helpers should be evaluated more positively than punishers and receive more help in the absence of any punishment threat.

Methods

This study was carried out in June 2014 using the online labour market, Amazon Mechanical Turk (hereafter MTurk; www.mturk.com). MTurk subjects represent a sample that is more diverse than the typical WEIRD ('Western, Educated, Industrialised, Rich, Democratic', Henrich et al. 2010) population used for most laboratory behavioural studies and than most other internet recruited populations (Buhrmester et al. 2011). Moreover, data collected via MTurk has been validated previously for running experimental economics studies: the findings are not qualitatively different

to those that are obtained under traditional laboratory settings (e.g. Mason and Suri 2012; Paolacci et al. 2010; Horton et al. 2011) even though low stakes are often used (Amir et al. 2012; Rand et al. 2012; Raihani et al. 2013). While there is evidence that experienced workers on MTurk may respond differently in games involving intuitive inference (e.g. Rand et al. 2014), a more recent article conducted in the same year (i.e. implying similar levels of experience among workers) has documented both internal and external validity of economic games on MTurk, showing that patterns of behavior are stable across different games and over time within the platform and - importantly - that helping behavior on MTurk correlates with real helping in a non-game setting (Peysakhovich et al. 2014).

We recruited 5241 US-based MTurk workers (2374 females; 2270 males; 597 did not specify; mean age = 30.2 ± 0.2; range = 18-73) and allocated them to the role of 'dictator' (n = 1004), 'receiver' (n = 1004), 'third-party' (n = 1421) or 'bystander' (n = 1810) in one of three treatments (described below). Note that the neutral terms Players 1 - 4 (respectively) were used in experimental instructions (see ESM). Prior to taking part in the study, all subjects were provided information about the task and were required to answer correctly three comprehension questions regarding the game structure. Subjects that failed one or more of the comprehension questions were not allowed to continue with the task and decisions from these individuals were therefore not recorded. An example of the text (excluding images) shown to subjects is available online (see ESM).

In each treatment, dictators were endowed with $0.50 and asked to choose between two options: (1) keep $0.25 and give $0.25 to receiver (hereafter the 'fair' option); or (2) keep $0.45 and give $0.05 to receiver (hereafter the 'selfish' option). Thus, dictators and receivers were involved in a dictator game (Kahneman et al. 1986), albeit with restrictions on the way that dictators could choose to split the endowment. Third-parties were allocated to one of the three treatments (below) and presented with either a 'selfish' or a 'fair' decision made by the dictator (see Table S1 for sample sizes). Third-

parties were given a starting endowment of $0.55 and asked to make a choice, depending on the treatment they were allocated to:

PUN treatment: Third-party could pay $0.05 to reduce dictator's bonus by $0.20 or do nothing (no cost).

HELP treatment: Third-party could pay $0.05 to increase receiver's bonus by $0.20 or do nothing (no cost).

PUN / HELP treatment: Third-party could pay $0.05 to reduce dictator's bonus by $0.20 or to increase receiver's bonus by $0.20, or do nothing (no cost).

Finally, bystanders were allocated to one of the three treatments above and presented with the dictator's decision (fair / selfish) and the third-party's response. Bystanders had a starting endowment of $1.05 and could choose to 'reward' the third-party (phrased as 'increasing Player 3's bonus' in the instructions) or do nothing. Rewarding the third-party cost the bystander $0.05 and increased the third-party's bonus by $0.25, whereas doing nothing entailed no cost to the bystander. The starting payoffs and fee-to-impact ratios of punishment and rewards available to third-parties and bystanders were chosen so as to ensure that disadvantageous inequity aversion could be ruled out as a possible motive for (i) third-parties reducing the dictator's bonus or (ii) bystanders not rewarding third-parties, respectively. Subjects in the dictator and third-party roles were not informed of the future possibility of being rewarded or punished by others when making their decisions in this game.

After subjects made their decision, they were asked to give a justification for the decision they made. These responses were free-form text to ensure that as much variation in responses could be captured as possible and also to avoid prompting people for answers they would not otherwise have given. We used bystanders' responses to ask whether the third-party's behavior was influential in the

decision to reward (or not reward) that individual. Specifically, third-party behaviour was deemed to be influential in the bystander's decision to reward / not reward if the bystander mentioned that their decision was influenced by the third-party's behaviour in the written text. Otherwise, we deemed that the third-party's behaviour had not been influential in the bystander's reward / not reward decision.

The written justifications also served as an additional comprehension check and subjects who gave answers that indicated they had not understood the game or the options available to them were excluded from the analyses. In addition, the responses from subjects who stated that they didn't believe the other players existed were also excluded from the analysis. In total, 41 responses from subjects allocated to the third-party role and 78 responses from subjects allocated to the bystander role were excluded, leaving total sample sizes of 1380 and 1732 for third-parties and bystanders, respectively, for analysis**.** All data, including responses that were ultimately excluded, are available as online supplementary material.

Our central questions were (i) whether third-party punishers would receive voluntary rewards from bystanders; and (ii) whether rewards were preferentially given to helpful rather than punitive third-parties when both options were available to the third-party. As such, in the paper, only the decisions made by bystanders were analysed, although the third-party decisions and accompanying analyses are presented in the ESM (Table S1).

Subjects made their decision in isolation, where third-parties were matched to the decision of dictators ex-post, and bystanders matched with the decisions of third parties ex-post (as in Raihani and McAuliffe 2012; Rand 2012)**.** For example, each of the 1421 third-parties was told that the dictator had either been fair or selfish but was actually only matched to a dictator who made the relevant decision after the experiment had finished. Similarly, bystanders were told of the dictator's

decision and the third-party's response but were only matched with individuals who made these decisions afterwards. As a result, the number of third-parties who chose to punish / help / do nothing in each treatment (Table S1) differ from the number of bystander decisions we recorded in response to these actions (Table 1).

Statistical Information

All data were analysed using R version 3.0.3 (www.r-project.org) using Chi-squared tests or Fisher's exact tests as appropriate. For the bystander decisions in each experimental treatment we produced a complete table of all pairwise comparisons. To correct for the increased risk of Type I errors (while maintaining appropriate power to reduce the risk of Type II errors) we used the Benjamini-Hochberg method (Benjamini and Hochberg 1995) to generate adjusted threshold levels for significance (α) as recommended by Waite and Campbell (2006). *P* values and adjusted α threshold levels for significance are reported in the tables of results. Agresti-Coull confidence intervals (95 %) for proportions (presented in figures) were calculated using the binom package in R (Dorai-Raj 2014).

Results

PUN treatment

In the PUN treatment (where punishment was the only option available to third-parties), the decision made by the third-party influenced the probability of receiving a reward from the bystander (Chi-squared test: $\chi^2 = 36.8$, $df = 3$, $P < 0.001$; Table 1). Third-parties that punished selfish dictators were more likely to be rewarded (99 / 162; 61.1 %) than individuals that did nothing in response to selfish dictators (78 / 157; 49.7 %), although this difference was only marginally significant at conventional levels and did not meet the Benjamini-Hochberg threshold level for significance (Chi-squared test: $\chi^2 = 3.77$, $df = 1$, $P = 0.052$, $\alpha = 0.042$; Figure 1a; Table 2). Nevertheless, third-parties who took no action in response to a selfish dictator were less likely to be rewarded than those who

took no action when the dictator was fair (Chi-squared test: $\chi^2 = 8.07$, $df = 1$, $P = 0.004$, $\alpha = 0.03$; Table 2), indicating that a third-parties were less likely to be rewarded for inaction when they were faced with a selfish dictator than when faced with a fair dictator. The context of punishment was important in predicting whether third-party punishers were rewarded for their actions. Players that punished a fair dictator were less likely to be rewarded (55 / 158; 34.8 %) than players that did nothing in response to a fair dictator (102 / 154; 66.2 %) (Chi-squared test: $\chi^2 = 29.6$, $df = 1$, $P < 0.001$; $\alpha = 0.008$; Figure 1a; Table 2).

HELP treatment

In the HELP treatment (where helping the receiver was the only option available to third-parties), the decision made by the third-party influenced the probability of receiving a reward from the bystander (Chi-squared test: $\chi^2 = 25.5$, $df = 3$, $P < 0.001$; Table 1). In response to selfish dictator allocations, 42 / 51 (82.4 %) helpful third-parties were subsequently rewarded compared with 26 / 57 (45.6 %) individuals that took no action (Chi-squared test: $\chi^2 = 14.0$, $df = 1$, $P = 0.0002$, $\alpha = 0.008$; Table 3; Figure 1b). For fair dictator allocations, 46 / 61 (75.4 %) of helpful third-parties were rewarded compared with 28 / 59 (47.5 %) of third-parties that did not help the receiver (Chi-squared test: $\chi^2 = 8.77$, $df = 1$, $P = 0.003$, $\alpha = 0.03$; Table 3; Figure 1b). Thus, unlike the PUN treatment, helpful third-parties were equally likely to be rewarded regardless of whether they were responding to fair or selfish dictator allocations (Chi-squared test: $\chi^2 = 0.44$, $df = 1$, $P = 0.51$, $\alpha = 0.04$; Table 3; Figure 1b).

PUN/HELP treatment

In the PUN/HELP treatment (where third-parties could choose between punishing or helping), the decision made by the third-party influenced the probability of receiving a reward from the bystander (Chi-squared test: $\chi^2 = 116.5$, $df = 5$, $P < 0.001$; Table 1). Bystanders were more likely to reward third-parties that punished selfish dictators (86 / 134; 64.2 % rewarded) than third-parties who took

no action (65 / 141; 45.7 % rewarded) (Chi-squared test: χ^2 = 8.36, df = 1, P = 0.004, α = 0.03; Table 4; Figure 2) indicating that, in this treatment at least, punishers were rewarded for their behaviour. Bystanders were also more likely to reward third-parties who helped the receiver (when the dictator was selfish) than third-parties who did nothing (Chi-squared test: χ^2 = 36.4, df = 1, P < 0.001, α = 0.02; Table 4; Figure 2). Bystanders apparently preferred to reward individuals that helped receivers over those that punished selfish dictators (rewards given to 86 / 134, 64.2 %, third-party punishers compared with 124 / 154, 79.4 %, third-party helpers; Chi-squared test: χ^2 = 8.88, *df* = 1, P = 0.002, α = 0.03; Table 4; Figure 2). As with the HELP treatment, third-party helpers were equally likely to be rewarded, regardless of whether the dictator was fair or not (Chi-squared test: χ^2 = 0.21, *df* = 1, P = 0.64, α = 0.05; Table 4; Figure 2).

Bystanders' justification of decisions

We summed the total number of times the bystander cited the third-party's action (punish the dictator versus help the receiver) in their decision to reward or not reward the third-party across all three conditions. We did not explore whether bystanders cited inaction by third-parties as influential in their decision to reward / not reward since we were interested in directly comparing why third-parties that helped were rewarded more often than third-parties that punished - even when punishment was justified. The probability that the bystander cited third-party behaviour in their decision to reward that individual varied according to whether the third-party had punished or helped and the context in which they made this decision (Fisher's exact test: P < 0.001; Table 5). When bystanders mentioned the third-party's helpful actions, it was always when justifying the decision to reward that individual and never when the bystander was not rewarding the third-party, regardless of whether the dictator had behaved fairly or selfishly (Fisher's exact test: P = 1, α = 0.05; Table 6). These data show that when helping behaviour was explicitly mentioned by the bystander, it was always deemed worthy of reward. By contrast, when the punishment was aimed at a selfish dictator, bystanders cited the third party's punishment behaviour as decisive for receiving

the reward in only 44 / 83 (53.0 %) cases with the remaining 47 % who mentioned the punisher's behaviour indicating that punishment was the reason they did *not* reward the third-party. Thus, even when the target of punishment was a selfish individual, the third-party's punitive behaviour was often invoked as a reason for withholding the reward from that individual, whereas helpful actions were never invoked as a reason for withholding rewards (Fisher's exact test: $P < 0.001$, $\alpha = 0.03$; Table 5 & 6).

Discussion

We asked whether third-party punishment would improve an individual's reputation and hence increase the probability of receiving voluntary help from bystanders in the future. The results show that third-party punishers were indeed rewarded by bystanders in the absence of any information on whether the punisher would behave cooperatively themselves (although this effect was only significant at conventional levels in the PUN/HELP treatment). Nevertheless, third-party helpers were rewarded even more than third-party punishers. The fact that third-party punishers were rewarded by uninvolved bystanders provides the first evidence that seemingly 'altruistic' punishment could in principle evolve via indirect reciprocity (i.e. in a way that is not altruistic in fitness terms). However, we also present evidence for considerable heterogeneity among bystanders in whether justified punishment is deemed as worthy of reward or not. Moreover, although we predicted that punishers would be most likely to be rewarded when punishment was the only way for these individuals to signal cooperative intent (i.e. in the PUN treatment rather than in the PUN/HELP treatment), our data did not support this prediction. It is therefore possible that bystanders do not consider whether individuals could have signalled cooperative intent through positive actions when assessing third-party punishers. Further empirical tests will be important to verify that this is indeed the case.

Although our results show that third-party punishers were rewarded by bystanders in this artificial

laboratory setting, the significance of indirect reciprocity as a mechanism underpinning cooperation in real-world settings is not clear (Roberts 2008). Observability promotes cooperation in real-world settings (hinting that people are sensitive to the reputation consequences of their behaviour) (e.g. Soetevent 2005; Yoeli et al. 2013) but demonstrating that individuals have concern for reputation is not the same thing as demonstrating that indirect reciprocity underpins cooperative actions. Evidence for indirect reciprocity requires a demonstration that (uninvolved) bystanders pay to reward (or help) individuals with a cooperative reputation which, to our knowledge, has not been demonstrated in a real-world setting (despite claims to the contrary, i.e. Yoeli et al. 2013). Rather than investing in costly rewards, bystanders might preferentially select cooperative individuals for interactions and it has been shown that reputation-based partner choice might be a more efficient mechanism for promoting cooperation than indirect reciprocity (Sylwester and Roberts 2013) because it does not require the bystander to pay a cost to reward the cooperative partner but instead to make a self-serving choice to interact with that individual (e.g. Bshary et al. 2006). Importantly, both indirect reciprocity and partner choice based accounts of cooperation predict that individuals invest more when reputation is at stake. Our current study shows that third-party punishers are deemed worthy of reward but in this experimental setting rewarding was the only way for bystanders to signal their approval of the third-party punisher's actions. In real-world settings, third-parties that invest to help victims or punish cheats might instead benefit because they are more likely to be chosen as interaction partners, not because they are rewarded (Raihani and Bshary 2015). Clearly, more work is required to determine how third-party punishers (and helpers) are treated under more naturalistic conditions.

Related to this point, it might also be interesting to explore whether similar reputation benefits might favour the evolution of punitive (or other aggressive) strategies in non-human animals. Evidence for punishment in non-humans is relatively rare (see Raihani et al. 2012 for a review) and third-party punishment has only been documented in one species that we are aware of (the cleaner

fish, *Labroides dimidiatus*). In the cleaner fish example, males punish females if the female cheating (biting the client) during a joint client inspection causes the client to leave (Raihani et al. 2010; 2011). Even though the client is the primary victim of the cheating female, the male is also disadvantaged by the female's behaviour and, crucially, benefits from investing in punishment because the female is more cooperative in future, leading to more productive encounters with clients for the male (Raihani et al. 2010). Male punishment of females is therefore self-serving which, according to our earlier predictions (Raihani and Bshary 2015), might make it unlikely that males would derive additional, reputation-based benefits from punishing a cheating female. Nevertheless, previous work on this system has shown that clients are sensitive to cleaner behaviour and preferentially interact with cooperative cleaners compared with cleaners of unknown cooperative level (Bshary et al. 2006) or who cheated (Pinto et al. 2011). It is possible, therefore, that clients who are cheated during a pairwise inspection will be more likely to return to that cleaning station if they see the male punishing the female than when no punishment is observed. In this hypothetical example, clients would make a self-serving decision to continue interactions at stations where service should improve and to terminate interactions at stations where service would continue to be poor. Thus, clients would be expected to prefer interactions with punitive male cleaners, even though male punishment of cheating females is self-serving rather than other-regarding. This experiment has not been conducted but would be an obvious setting to explore the reputation consequences of punishment in non-humans.

Our finding that third-party helpers were more likely than third-party punishers to be rewarded by bystanders might be one reason why individuals typically prefer to help rather than punish (e.g. Charness et al. 2008; Chavez and Bicchieri 2013; Lotz et al. 2011; see ESM for supportive analysis of third-party decisions in this game). There are also plausible alternative explanations for preferences to help victims rather than punish cheats, which might be more salient under real-world settings. The vast majority of studies investigating the conditions under which third-parties will

invest to punish others have been performed in laboratory settings (e.g. Charness et al. 2008; Fehr and Fischbacher 2004; Kurzban 2007; Piazza and Bering 2008) whereas real-world studies of the conditions predicting third-party punishment - and the consequences of such investments - are scarce. Although retaliation was not a component of our study, in real-world settings, players that punish might be more likely to be victims of counter-punishment (e.g. Dreber et al. 2008; Herrmann et al. 2008) than individuals that do nothing or that help others instead. The threat of retaliation from the target (and associated costs incurred by the punisher) might lead individuals to avoid punishment when other non-confrontational options are available. Another reason why humans might generally prefer to help victims rather than punish cheats is because, under real-world scenarios, the act of helping an individual in need might create opportunities for direct reciprocity from that individual or mean that the helpful individual is more likely to be chosen for subsequent interactions (either by the recipient of the help or by an uninvolved bystander) (e.g. see Adams and Mullen 2013; Sylwester and Roberts 2013). These possibilities all remain open for empirical testing.

We note that the current findings seem to contrast with some earlier work done by Kiyonari and Barclay (2008), who showed that individuals that invested in the punishment of uncooperative group members were not more likely to be rewarded than non-punishers. In line with our study, however, individuals that rewarded cooperative group members were rewarded by others. We note that a key difference between this study and Kiyonari and Barclay (2008) is that, in our setup, the punishers were themselves not involved in the initial dictator game, whereas in Kiyonari and Barclay (2008) the punishers were also disadvantaged by the presence of uncooperative individuals in the group and stood to benefit from changing the behavior of these defectors via punishment. Individuals that stand to derive any self-serving benefit from their punishment are expected to be viewed differently to individuals who apparently invest with no expectation of recouping this investment in the future: the latter may be viewed by bystanders as having pure other-regarding

preferences whereas the former might be interpreted as more competitive and therefore less deserving of rewards (Raihani and Bshary 2015). We further note that although third-party punishers were more likely than non-punishers to be rewarded by bystanders, the proclivity to reward punitive individuals was not universal in our sample. Indeed, many of the bystanders cited the punisher's behavior as a key factor underpinning their decision *not* to reward, even when the punishment was aimed at a selfish target. We do not know what factors underpin this heterogeneity in responses to punishers, although this would be a useful avenue for further exploration if we are to understand whether and how third-party punishers benefit from a punitive reputation. In contrast to punishment, the written text justifying the decision to reward the third-party showed that helping behavior was always viewed as positive.

The results presented here also have implications for the decision rules used by individuals when deciding who to help. Explanations for the evolution of indirect reciprocity differ in how donors decide whether recipients are good (and therefore deserving of help) or bad (and therefore undeserving of help) (Ohtsuki and Iwasa 2004) The simplest strategies require donors to judge recipients based on the recipient's previous behavior: recipients that cooperated previously receive help whereas recipients that defected previously do not. Despite their simplicity, it has been shown these so-called 'scoring' strategies (Nowak and Sigmund 1998a) are not evolutionarily stable since donors that refuse to help bad recipients themselves acquire a bad reputation and so are refused help in the future (Leimar and Hammerstein 2001). Instead, strategies where the donor assesses the 'standing' (Sugden 1986) of the recipient before deciding whether to help tend to outperform scoring strategies in evolutionary simulations and, unlike scoring strategies, do not depend on high levels of genetic drift and low costs of helping in small populations (Leimar and Hammerstein 2001; Ohtsuki and Iwasa 2004). Unlike scoring strategies, donors do not risk harming their own reputation if they justifiably refuse to help a recipient in bad standing. The most successful strategies favouring the evolution of indirect reciprocity tend to insist that recipients in good standing are always helped; but

also approve of defection against recipients in bad standing (Ohtsuki and Iwasa 2004). Despite the theoretical dominance of standing strategies, there is little empirical evidence that humans actually use them (Ule et al. 2009) perhaps, it has been argued, because second and third-order moral assessment rules are too cognitively demanding (Milinski et al. 2001). Our results offer some evidence that individuals can use third-order assessment rules when deciding who to help, at least in the context of punishment, since unjustified punishment (aimed at fair dictators) was rewarded less often than justified punishment (though we note that the third-order assessment rules in our one-shot game were also likely to have been cognitively less demanding than those that players would have had to use in the multi-round game used by Milinski et al. 2001). It may be that context is especially important when considering how to interpret punitive actions by others but that helping behavior is more often viewed as positive, regardless of whether it is justified or not. It is difficult to judge from our study whether individuals are more likely to use a scoring or a standing strategy when judging decisions to help or not since third-parties were directing help at receivers, who had no known reputation. Thus, bystanders could not condition their response to third-party behavior on the basis of how receivers behaved in this study. This may be why we observed that helpful players were rewarded more than unhelpful players, without any additional effect of the context. In a future study, it would be interesting to ask how bystanders would have been treated by another bystander (i.e. a fifth player) in a subsequent round. For example, would bystanders that refused to reward unjustified third-party punishment be more or less likely to be helped by future partners? Under a scoring strategy we expect them to be refused help; but they should be helped if others are using a standing strategy.

It is unlikely that the patterns we observe arise as an artefact of either using low stakes in an online setting or because both third-parties and bystanders could confer a benefit on another player for a relatively small cost to themselves ($0.05). Several previous studies have shown that patterns of behavior on MTurk, using low stakes, reliably match those that are seen in more traditional

laboratory settings using higher stakes (e.g. Suri and Watts 2011; Horton et al. 2011; Amir et al. 2012; Rand et al. 2012; Raihani et al. 2013). Moreover, the fee-to-impact ratio of 1:4 for third-parties and 1:5 for bystanders is comparable to that which has been used in laboratory studies investigating punishment and reward (e.g. Nikiforakis et al. 2007; Nikiforakis & Normann 2008; Dreber et al. 2008; Horita 2010; Ozono and Watabe 2012), with some colleagues even investigating scenarios where punishment and reward are costless (Cushman et al. 2009). Most importantly, neither low stakes nor fee-to-impact ratio can explain why subjects would respond differently to third parties who had either punished selfish individuals, or helped victims, or done nothing. As pointed out by Kummerli et al (2010), experimental studies are most informative with respect to the question how a variable of interest affects behaviour in a qualitative way, while precise measures of resulting payoffs may often be misleading. Thus, the propensity to reward punishers in our experiment is a relative, rather than an absolute, measure of behaviour.

In conclusion, we have shown that, so long as punishment is justified, third-party punishers are more likely to be rewarded for their actions than individuals who take no action against selfish others. However, helpful individuals are rewarded even more than punishers and we saw higher levels of agreement among bystanders on the merits of helping behavior than on the merits of punishment. We hope that these results will stimulate theoreticians to develop models of the evolution of punishment based on indirect reciprocity and stimulate empiricists to explore in more detail when and why punishers might sometimes be loved.

ACKNOWLEDGMENTS

We would like to thank Katie McAuliffe for useful discussion in the early stages of analysis. This work was funded by a Royal Society University Research Fellowship to NR.

LITERATURE CITED

Adams G.S., and E. Mullen. 2013. Increased voting for candidates who compensate victims rather than punish offenders. Soc. Jus. Res. 26: 168–192.

Amir O., D.G. Rand, and Y.G. Gal. 2012. Economic games on the internet: the effect of $1 stakes. PLoS ONE 7: e31461.

Barclay P. 2006. Reputational benefits for altruistic punishment. Evol. Hum. Behav. 27: 325–344.

Benard, S. 2013. Reputation systems, aggression, and deterrence in social interaction. Soc. Sci. Res. 42: 230-245.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57: 289-300.

Bone, J. and N.J. Raihani (in press) Human punishment is motivated by both a desire for revenge and a desire for equality. Evol. Hum. Behav.

Brandt H., C. Hauert, and Sigmund K. 2003. Punishment and reputation in spatial public goods games. Proc. R. Soc. B 270: 1099–1104.

Bshary R., and A.S. Grutter. 2006. Image scoring causes cooperation in a cleaning mutualism. Nature 441, 975-978.

Bshary R., and J.L. Bronstein. 2011. A general scheme to predict partner control mechanisms in pairwise cooperative interactions between unrelated individuals Ethology 117: 1-13.

Buhrmester M., T. Kwang, and S.D. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Persp. Psych. Sci. 6: 3–5.

Charness G., R. Cobo-Reyes, and N. Jiminez. 2008. An investment game with third-party intervention. J. Econ. Behav. Org. 68: 18–28.

Chavez A.K., and C. Bicchieri. 2013. Third-party sanctioning and compensation behavior: Findings from the ultimatum game. J. Econ. Psychol. 39: 268–277.

Clutton-Brock, T.H., and G.A. Parker. 1995. Punishment in animal societies. Nature 373: 209–216.

Cushman F., A. Dreber, Y. Wang, and J. Costa. 2009. Accidental outcomes guide punishment in a 'trembling hand' game. PLoS ONE 4: e6699.

Dorai-Raj S. 2014. binom: Binomial Confidence Intervals For Several Parameterizations. R package version 1.1-1. http://CRAN.R-project.org/package=binom

dos Santos M., D.J. Rankin, and C. Wedekind. 2011. The evolution of punishment through reputation. Proc. R. Soc. B 278: 371–377.

dos Santos M., D.J. Rankin, and C. Wedekind. 2013. HUMAN COOPERATION BASED ON PUNISHMENT REPUTATION. Evolution 67: 2446–2450.

Dreber A., D.G. Rand, D. Fudenberg, and M.A. Nowak. 2008. Winners don't punish. Nature 452: 348–351.

Fehr E., and U. Fischbacher. 2004. Third-party punishment and social norms. Evol. Hum. Behav. 25: 63–87.

Fehr E., and S. Gächter. 2002. Altruistic punishment in humans. Nature 415: 137–140.

Gordon, D.S., J.R. Madden, and S.E.G. Lea. 2014. Both loved and feared: third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. PloS ONE 9, e110045.

Henrich J., S.J. Heine, and A. Norenzayan. 2010. The weirdest people in the world? Behav. Brain Sci. 33: 61–83.

Herrmann B., C. Thöni, and S. Gächter. 2008. Antisocial punishment across societies. Science 319: 1362–1367.

Hilbe C., and K. Sigmund. 2010. Incentives and opportunism: from the carrot to the stick. Proc. R. Soc. B 277: 2427-2433.

Hilbe C., and A. Traulsen. 2012. Emergence of responsible sanctions without second-order free riders, antisocial punishment or spite. Sci. Rep. 458:srep00458.

Horita Y. 2010. Punishers May Be Chosen as Providers But Not as Recipients. Lett. Evol. Behav. Sci. 1: 6-9.

Horton J.J., D.G. Rand, and R.J. Zeckhauser. 2011. The online laboratory: conducting experiments in a real labor market. Exp. Econ. 14: 399–425.

Kahneman D., J.L. Knetsch, and R. Thaler. 1986. Fairness as a constraint on profit seeking: entitlements in the market. Am. Econ. Rev. 76: 728-741.

Kiyonari T., and P. Barclay. 2008. Cooperation in social dilemmas: free-riding may be thwarted by second-order reward rather than by punishment. J. Pers. Soc. Psychol. 95: 826-842.

Kummerli, R., M.N. Burton-Chellew, A. Ross-Gillespie, and S.A. West. 2010. Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games. Proc. Natl. Acad. Sci. USA 107: 10125-10130.

Kurzban R., P. Descioli, and E. O'Brien. 2007. Audience effects on moralistic punishment. Evol. Hum. Behav. 28: 75–84.

Leimar O., and P. Hammerstein. 200.1 Evolution of cooperation through indirect reciprocity. Proc. R. Soc. B 268:745-753.

Lotz S., T.G. Okimoto, T. Schlosser, and D. Fetchenhauer. 2011. Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. J. Exp. Soc. Psychol. 47: 477–480.

Mason W., and S. Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. Behav. Res. Meth. 44: 1–23.

Milinski M., D. Semmann, T. Bakker, and H. Krambeck. 2001. Cooperation through indirect reciprocity: image scoring or standing strategy? Proc. R. Soc. B 268: 2495-2501.

Nelissen R. 2008 The price you pay: cost-dependent reputation effects of altruistic punishment. Evol. Hum. Behav. 29: 242–248.

Nikiforakis N., and H-T. Normann. 2008. A comparative statics analysis of punishment in public-good experiments. Exp. Econ. 11: 358-369.

Nikiforakis N., H-T. Normann, and B. Wallace. 2010. Asymmetric enforcement of cooperation in a social dilemma. South. Econ. J. 76: 638-659.

Nowak M.A., and K. Sigmund. 1998a. The Dynamics of Indirect Reciprocity. J. Theor. Biol. 194:

Nowak M.A., and K. Sigmund. 1998b. Evolution of indirect reciprocity by image scoring. Nature

393:573–577.

Ohtsuki H., and Y. Iwasa . 2004. How should we define goodness? - reputation dynamics in indirect reciprocity. J. Theor. Biol. 231:107–120.

Ohtsuki H., Y. Iwasa and M.A. Nowak. 2009. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature, 457: 79–82.

Ozono H,, and M. Watabe. 2012. Reputational benefit of punishment: comparison among the punisher, rewarder, and non-sanctioner. Lett. Evol. Behav. Sci. 3: 21–24.

Paolacci G., J. Chandler, and P.G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. Judg. Dec. Mak. 5:411–419.

Pedersen E.J., R. Kurzban, and M.E. McCullough. 2013. Do humans really punish altruistically? A closer look. Proc. R Soc. B 280:20122723.

Peysakhovich A., M.A. Nowak MA, and D.G. Rand. 2014. Humans display a 'cooperative phenotype' that is domain general and temporally stable Nat. Comm. 5: 4939.

Piazza J., and J. Bering. 2008. The effects of perceived anonymity on altruistic punishment. Evol. Psychol. 6: 487–501.

Pinto, A., J. Oates, A. Grutter and R. Bshary. 2011. Cleaner wrasses *Labroides dimidiatus* are more cooperative in the presence of an audience. Curr. Biol. 21: 1140-1144.

Raihani N.J., and R. Bshary. 2015. The reputation of punishers. Trends Ecol. Evol. 30: 98-103.

Raihani N.J., and K. McAuliffe. 2012. Human punishment is motivated by inequity aversion, not a desire for reciprocity. Biol. Lett. 8: 802–804.

Raihani N.J., A. Thornton, and R. Bshary. 2012. Punishment and cooperation in nature. Trends Ecol. Evol. 27: 288–295.

Raihani N.J., R. Mace, and S. Lamba. 2013. The effect of $1, $5 and $10 stakes in an online dictator game. PLoS ONE 8: e73131.

Rand D.G. 2012. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. J. Theor. Biol. 299: 172–179.

Rand D.G., J.D. Greene, and M.A. Nowak. 2012. Spontaneous giving and calculated greed. Nature 489: 427-430.

Rand D.G., A. Peysakhovich, G.T. Kraft-Todd, G.E. Newman, O. Wurzbacher, M.A. Nowak, and J.D. Greene. 2014. Social heuristics shape intuitive cooperation. Natt. Comm. 5: 3677.

Roberts, G. 2008. Evolution of direct and indirect reciprocity. Proc. R Soc. B 275:173-179.

Rockenbach B., and M. Milinski. 2011. To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. Proc. Natl. Acad. Sci. USA 108: 18307–18312.

Roos P., M. Gelfand, D. Nau, and R. Carr. 2014. High strength-of-ties and low mobility enable the evolution of third-party punishment. Proc. R. Soc. B 281: 20132661.

Seinen I., and A. Schram. 2006. Social status and group norms: Indirect reciprocity in a repeated helping experiment. Eur. Econ. Rev. 50:581–602.

Sigmund K., C. Hauert, and M.A, Nowak. 2001. Reward and punishment. Proc Natl Acad Sci. USA 98: 10757-10762.

Sugden R. 1986. The economics of rights, co-operation and welfare. Oxford, UK: Blackwell.

Suri S., and D.J. Watts. 2011. Cooperation and Contagion in Web-Based, Networked Public Goods Experiments. PLoS ONE 6: e16836.

Sylwester, K., and G. Roberts. 2013. Reputation-based partner choice is an efficient alternative to indirect reciprocity in solving social dilemmas. Evol. Hum. Behav. 34, 201-206.

Ule A., A. Schram, A. Riedl, and T.N. Cason. 2009. Indirect Punishment and Generosity Toward Strangers. Science 326:1701–1704.

Waite, T.A. and L.G. Campbell. 2006. Controlling the false discovery rate and increasing statistical power in ecological studies. Ecoscience 13: 439-442.

Wedekind, C. and M. Milinski. 2000. Cooperation through image scoring in humans. Science 288, 850-852.

Yoeli, E., M. Hoffmann, D.G. Rand, and M.A. Nowak. 2013. Powering up with indirect reciprocity

in a large-scale field experiment. Proc Natl Acad Sci. USA 110, 10424-10429.

Figure 1. The proportion of third-parties (P3) who were rewarded by bystanders following their decision to do nothing versus (a) punish the dictator (PUN Treatment) or (b) help the receiver (HELP Treatment). Dark bars that the dictator was fair; light bars indicate that the dictator was selfish. Bars represent Agresti-Coull confidence intervals (95 %).

Figure 2. The proportion of third-parties (P3) who were rewarded by bystanders according to their decision to either punish the dictator, help the receiver or do nothing (PUN / HELP Treatment). Dark bars indicate responses when dictator was fair; light bars show responses when dictator was selfish. Bars represent Agresti-Coull confidence intervals (95 %).

Table 1. Bystander decisions to reward (REW) or not reward (NO REW) the third-party, according to the treatment, the dictator's decision (fair / selfish) and the third-party's decision (P = punished dictator; H = helped receiver; NO = took no action).

Table 2. *P* values associated with pairwise comparisons for bystander decisions to reward third-parties in the PUN treatment. All comparisons were performed with Chi-squared tests. PUN and NO refer to third-parties who punished the dictator or took no action, respectively. F and S refer to whether the dictator was fair or selfish, respectively. The Benjamini-Hochberg method (Benjamini and Hochberg 1995) was used to adjust the critical threshold (α) for determining significance (thresholds presented below *P* values in parentheses). All results that significant at $P < \alpha$ are highlighted. The percentage of bystanders who rewarded third-parties in each of the conditions is

presented in parentheses.

Table 3. *P* values associated with pairwise comparisons for bystander decisions to reward third-parties in the HELP treatment. All comparisons were performed with Chi-squared tests. HELP and NO refer to third-parties who helped the receiver or took no action, respectively. F and S refer to whether the dictator was fair or selfish, respectively. Critical thresholds (α) for determining significance are presented below *P* values in parentheses. All results that significant at $P < \alpha$ are highlighted. The percentage of bystanders who rewarded third-parties in each of the conditions is presented in parentheses.

Table 4. *P* values associated with pairwise comparisons for bystander decisions to reward third-parties in the PUN/HELP treatment. All comparisons were performed with Chi-squared tests. PUN, HELP and NO refer to third-parties who punished the dictator, helped the receiver or took no action, respectively. F and S refer to whether the dictator was fair or selfish, respectively. Critical thresholds (α) for determining significance are presented below *P* values in parentheses. All results that significant at $P < \alpha$ are highlighted. The percentage of bystanders who rewarded third-parties in each of the conditions is presented in parentheses.

Table 5. Bystanders' who cited third-party's behaviour in the written text justifying their decision to reward or not reward the third-party. Bystander justifications were summed across the three experimental treatments.

Table 6. *P* values associated with pairwise comparisons for whether bystander cited third-party's behaviour when making decision to reward or not reward the third-party. Due to low cell counts for some cells, all comparisons were performed with Fisher's Exact tests tests. PUN and HELP refer to third-parties who punished the dictator or helped the receiver, respectively. F and S refer to whether the dictator was fair or selfish, respectively. Critical thresholds (α) for determining significance are

presented below *P* values in parentheses. All results that significant at $P < \alpha$ are highlighted. The percentage of bystanders who cited the third-party's behaviour in their justification to reward the third-party is presented in parentheses for each condition.