PROCEEDINGS OF THE ROYAL SOCIETY B

rspb.royalsocietypublishing.org

Research

Cite this article: Pedersen EJ, Kurzban R, McCullough ME. 2013 Do humans *really* punish altruistically? A closer look. Proc R Soc B 280: 20122723.
http://dx.doi.org/10.1098/rspb.2012.2723

Received: 16 November 2012
Accepted: 8 February 2013

Subject Areas:
behaviour, cognition, evolution

Keywords:
cooperation, altruistic punishment, third-party punishment, affective forecasting, evolutionary psychology

Author for correspondence:
Michael E. McCullough
e-mail: mikem@miami.edu

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2012.2723 or via http://rspb.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

Do humans *really* punish altruistically? A closer look

Eric J. Pedersen[1], Robert Kurzban[2,3] and Michael E. McCullough[1]

[1]Department of Psychology, University of Miami, Coral Gables, FL 33124-0751, USA
[2]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104-6241, USA
[3]Department of Economics, University of Alaska, Anchorage, AK 99508, USA

Some researchers have proposed that natural selection has given rise in humans to one or more adaptations for *altruistically* punishing on behalf of other individuals who have been treated unfairly, even when the punisher has no chance of benefiting via reciprocity or benefits to kin. However, empirical support for the altruistic punishment hypothesis depends on results from experiments that are vulnerable to potentially important experimental artefacts. Here, we searched for evidence of altruistic punishment in an experiment that precluded these artefacts. In so doing, we found that victims of unfairness punished transgressors, whereas witnesses of unfairness did not. Furthermore, witnesses' emotional reactions to unfairness were characterized by envy of the unfair individual's selfish gains rather than by moralistic anger towards the unfair behaviour. In a second experiment run independently in two separate samples, we found that previous evidence for altruistic punishment plausibly resulted from affective forecasting error—that is, limitations on humans' abilities to accurately simulate how they would feel in hypothetical situations. Together, these findings suggest that the case for altruistic punishment in humans—a view that has gained increasing attention in the biological and social sciences—has been overstated.

1. Introduction

In many animal species, including humans, individuals punish conspecifics that have harmed them [1–3]. Some researchers have recently argued that humans, unlike other animals, also *altruistically* punish individuals who have harmed others, even when the punisher has no chance of benefiting via reciprocity or benefits to kin [4–6]. Results from several economics experiments appear to support this claim [4,6,7], but some scholars have questioned both the adaptationist logic behind such theoretical claims [8–10] and the interpretation of the empirical results [8,10–14]. Here, we elide these theoretical debates and instead investigate a more basic empirical question: do people actually spontaneously punish individuals who have only harmed other individuals in anonymous settings in the laboratory? Put differently, do the empirical research findings often marshalled in support of the altruistic punishment hypothesis [4,6] provide a reliable guide to the presence or absence of a propensity for altruistic punishment in humans?

In previous work, researchers claimed empirical support for the existence of altruistic punishment on the basis of results from public goods game experiments in which the individual being punished had harmed—or failed to help—the putative punisher as well as other victims [4], leaving open the possibility that the punishment was vengeful, rather than altruistic [10]. Results from similar experiments that exclude revenge as a possible motive suggest that investments in punishment in such contexts are conspicuously low [15]. Additional data frequently adduced in support of the altruistic punishment hypothesis come from third-party punishment games [6,7], in which a *dictator* chooses to give some portion of a sum of money (or none) to a passive *recipient*. A third player can punish the dictator (at a cost) in response to the dictator's transfer to the recipient. Many third parties in games with this structure pay

a cost to punish stingy dictators, despite receiving no financial benefit from doing so [6,7].

However, five methodological limitations of the standard third-party punishment game might conspire to yield inflated estimates of humans' propensity to punish strangers for having behaved unfairly towards other strangers. First, in the standard game, subjects are assigned to a third-party role that implies their task is to determine how much to punish the dictator; indeed, the only choices third parties can make are whether to punish the dictator [16] (and, if so, how much). Thus, any error will lead to an increase in the estimated quantity of punishment. Second, punishment in the third-party punishment game is typically administered with the presence (or inferred presence) of an audience: punishment of the dictator by the third party is witnessed by the initial victim because all players see the results of the game. The presence of an audience introduces reputational considerations that could motivate punishment as a means of pursuing indirect fitness benefits (e.g. by signalling one's quality as a cooperative partner [17,18], or one's formidability to prevent future exploitation of oneself [19,20] or one's friends and kin [21]). Indeed, it has been shown—though with a different paradigm—that observers of unfair treatment punish third parties significantly less when they are assured no one will see their decision [22] (however, see [23]).

Third, the third-party punishment game is typically conducted with the 'strategy method' [24], which requires third parties to repeatedly respond to a series of hypothetical dictator choices—in advance of learning of the dictator's actual choice—that are progressively more (or progressively less) unfair [6]. Such methods can cause subjects to infer that the experimenters expect them to vary their responses according to some feature that varies across the set of repeated scenarios [25]. Consequently, owing to a well-known experimental artefact called *demand*, subjects might feel compelled to punish at least some of the time, calibrating those decisions to the only feature of the dictators' repeated choices that varies: how unfair they are. This is especially problematic in the standard third-party punishment game because rewarding is not allowed; the only way subjects can vary their responses is to vary their amount of punishment. In a notable exception, Almenberg *et al*. [26] did add a rewarding option to the typical third-party punishment game (conducted with the strategy method), and a small amount of third-party punishment was observed, on average, when dictators transferred \$0 (of \$10) to the recipient. We note, however, that subjects in this experiment were informed, before making their decisions, that it was possible they would not be paired with another subject—in such a case, their decisions would not be enacted and they would retain all of their money (i.e. participants' decisions were somewhat hypothetical; see below).

Fourth, the strategy method also involves affective forecasting [27] in as much as it requires subjects to respond *ex ante* to dictator actions that have not yet occurred. Such behavioural commitments can differ from the actual behaviours people enact after experiencing social situations directly because people frequently weight the features of social situations differently during conscious deliberation than they do after experiencing those social situations in real time [28]. For example, as forecasters, people severely overestimate how upset they would feel by (and subsequently, how much they would attempt to avoid interacting with) someone who had made a racist comment; by contrast, subjects who have actually observed another individual express strongly racist attitudes (versus those in control conditions) respond with relative indifference to the racist individual [29].

Fifth, previous claims that anger is the predominant emotional response of third-party punishers have relied on self-reports of anger in response to hypothetical scenarios [4,6]. Self-reports of anger are typically highly correlated with self-reports of other, similar emotions—including envy [30]. To the extent that the covariation between self-reported anger and self-reported envy is not statistically controlled, estimates of third parties' anger towards unfair strangers might actually reflect envy, which can also motivate costly punishment in pursuit of goals that are quite distinct from putatively altruistic goals such as enforcing norms or delivering deterrence benefits to strangers [31]. Specifically, if third parties' punishment of individuals who have treated another individual unfairly is motivated by envy, but not by anger, then the mechanisms that motivate third-party punishment might process cues that another individual has obtained better outcomes than the self, rather than cues that an individual has violated a norm or harmed an anonymous third party in whom the punisher has no fitness interest [6,7].

Here, we present two experiments designed to test whether subjects punish altruistically on behalf of strangers in a third-party punishment game that was designed to rectify the methodological problems noted above. We also examined whether previous findings could plausibly be explained as a product of affective forecasting errors. We note that our goal was not to estimate the unique influence of each of these five potential methodological problems; rather, our goal was to test whether the altruistic punishment hypothesis could survive falsification in an experiment that eliminated these problems. Experiment 1 was a modified third-party punishment game in which subjects could either punish *or* reward—thereby reducing experimental demand for punishment [25], the confounding of error and punishment, and potential audience effects. Also, subjects made decisions about giving or deducting money from dictators *after* witnessing the dictator's decision, which enabled us to measure third-party punishment without the possibility of affective forecasting errors [27]. Additionally, our measures of emotion were fine-grained enough that it was possible to evaluate the unique motivational roles of anger and envy. In experiments 2a and 2b, the same third-party punishment game was presented as a hypothetical vignette to subjects from two different research pools.

2. Methods

(a) Subjects

(i) Experiment 1

Subjects were 315 University of Miami undergraduates (mean age = 19.12, s.d. = 2.99; 57% female). They received partial course credit and monetary compensation (see below).

(ii) Experiment 2a

Subjects were 538 individuals (mean age = 34.37, s.d. = 12.14; 60% female) recruited via Amazon Mechanical Turk (http://www.mturk.com/mturk/welcome) and were paid \$0.25 for their participation. Participation was restricted to users in the

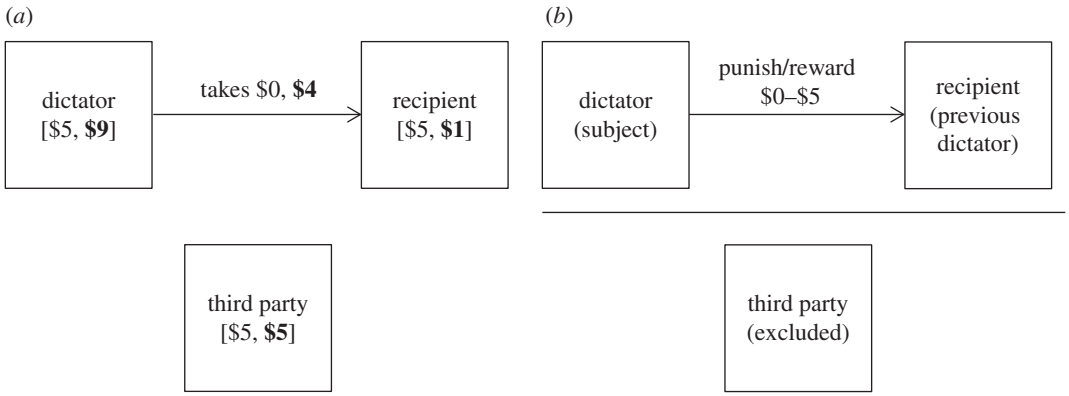

Figure 1. Game structure. (*a*) Round 1: all players started with $5. Subject was either recipient or third party. The dictator (a fictitious player whose 'decisions' were determined by computer script) either took $0 (fair conditions; unbolded) or $4 (unfair conditions; bolded) from recipient. (*b*) Round 2: Players started with $5. Subject was dictator; previous 'dictator' was recipient; third party was excluded. Subject was allowed to give any portion of $5, do nothing or pay a 1 : 4 cost to deduct money from recipient. Money deducted from recipient in round 2 was 'burned'—it was not gained by dictators as income.

USA. Because participants merely had to read a vignette and then report their forecasts of how they would think, feel and behave if the hypothetical situation had actually happened to them, participation generally took about 4 min.

(iii) Experiment 2b

We replicated experiment 2a with University of Miami undergraduates; 394 subjects (mean age = 18.74, s.d. = 1.27; 53% female) participated for partial course credit.

(b) Procedure

(i) Experiment 1

Subjects were run in individual sessions at a computer station in an isolated room (see the electronic supplementary material, S2.1). The entire experiment, including instructions, was conducted on a computer via E-Run with a script created in E-Prime v. 2.0. After subjects provided informed consent, they were told they would be interacting with two other players in the building over the computer network and that it was important that those other people remain anonymous; in fact, they interacted with a pre-programmed computer script. Without this deception, the research would have been unfeasible (see the electronic supplementary material, S2.1). Subjects were informed that they would be participating in an economic decision-making game that would last for multiple rounds and they would be paid based on the money they earned during the game. Because deception was involved, everyone was paid a flat rate of $9 at the end of the experiment following a debriefing. We used a 'funnel debriefing' method designed to detect suspicion and explore subjects' reactions to having been deceived [32]. Subjects flagged for suspicion were excluded from all analyses presented; re-including them in analyses did not qualitatively affect the results in any way (see the electronic supplementary material, S1.2).

The decision-making game comprised two rounds (figure 1) in which each player was given $5 to use in each round and assigned to one of three roles: decision-maker, receiver or observer. (We refer to these roles here as *dictator*, *recipient* and *third party*, respectively, to be consistent with labels used in previous work on third-party punishment.) Subjects were not told the exact number of rounds, to avoid end of game effects [33], and were told that money earned during each round would be 'banked' and thus unaffected by subjects' behaviour during subsequent rounds. The dictator ostensibly had the option to give any portion of his or her $5 to the recipient, or take any portion of the recipient's $5; the third party would merely see the results of the round and would not be affected by the dictator's choice. Subjects were informed that in some rounds all players would be involved, and in other rounds some players might be excluded. Subjects were randomly assigned to be either the third party or the recipient in the first round, and the (computer-programmed) dictator either took $4 or $0 from the recipient. The computer displayed a summary screen for the round showing the amount of money each player earned for the round. Following the round, subjects completed a lexical decision task (see the electronic supplementary material, S1.3) and a series of self-report questions (see below).

Prior to role assignment for the second round, subjects were informed that there would be no third party in round 2 (to avoid potential audience effects [22,34]); one player would be assigned to a different task and be unable to see the results of the interaction. We note that the presence of the experimenter can also induce audience effects [22]; we took great care to minimize this potential influence by (i) clearly informing participants during the consent process that their data would be stored completely anonymously and could not be connected to them in any way, and (ii) minimizing contact with the experimenter by presenting all instructions electronically. Although we cannot rule out experimenter audience effects completely, our results are as insulated from them as we believe was possible in the context of this experiment.

All players were given another $5; because previous earnings had been 'banked', all players started round 2 with $5. The subject was assigned the role of dictator while the dictator from round 1 (who had treated either the subject or the other player fairly or unfairly) was assigned the role of the recipient (ostensibly by chance). Players were identified consistently throughout, so subjects were aware that the recipient in round 2 was the same player that had been the dictator in round 1. Subjects were instructed that they could give any amount of their $5 to the recipient, do nothing, or remove any amount of the recipient's $5 (the word 'punishment' was never used). Removing money cost one-quarter of the amount removed and, unlike in the first round, was not gained by the subject as income—it simply disappeared. Note that the cost of punishment used here, 1 : 4, was less expensive than the 1 : 3 cost typically used in the third-party punishment game; previous research has shown that punishment becomes more likely as the cost of punishment declines (see [10] for review). Following the completion of the round, the experiment ended and the experimenter debriefed the subject through an extensive, staged process to assess the believability of the experiment and to explain why deception was necessary [32].

(ii) Experiments 2a and 2b

After providing consent, subjects were instructed to imagine themselves 'in a particular situation in our laboratory. Please

try to picture yourself in the situation we are describing. We will ask you to complete a series of questions regarding how you think you would think, feel and act in this situation.' The layout and instructions of the game were presented as they were in experiment 1, and the rounds of the game and the self-report measures were the same. However, subjects did not complete a lexical decision task following the first round and were not debriefed following the completion of the experiment (because no deception was involved).

(c) Psychometric information regarding the self-report measures

(i) Self-rated emotions towards the other players

Subjects were asked to describe their emotional responses towards both of the other players after the first round. They described their feelings towards both players to avoid demand effects that might have occurred by probing only about the dictator. (Emotional reactions to the other player were not of theoretical interest here and so, in the interest of brevity, we do not report them.)

— *Anger*. Three-item composite of ratings on a scale from 0 (*not at all*) to 5 (*extremely*) of the extent to which the subject was 'angry,' 'mad,' and 'outraged' at the dictator (Cronbach's $\alpha = 0.94$).
— *Envy*. Two-item composite of ratings on a scale from 0 (*not at all*) to 5 (*extremely*) of the extent to which the subject was 'envious' and 'jealous' of the dictator (Cronbach's $\alpha = 0.84$).

(ii) Fairness/moral wrongness of the round 1 dictator's behaviour

Subjects were asked to rate both how 'fair' and how 'morally wrong' the dictator's behaviour was towards the recipient in round 1 on a scale from 1 (*not at all*) to 9 (*totally*).

3. Results

(a) Experiment 1

Third parties did not punish on behalf of strangers: a one-sample Wilcoxon test (used because distributions were non-normal) revealed that the sample median of the distribution of dollars punished or rewarded in round 2 (in terms of the effect on the recipient, not the cost to the subject) by third-party witnesses of unfairness did not differ significantly from a hypothesized median of zero ($z = -1.48$, $p = 0.140$, $n = 65$; all p-values throughout manuscript are two-tailed). By contrast, victims of unfairness punished a non-zero amount ($z = -3.52$, $p < 0.001$, $n = 61$)—significantly more than mere witnesses of unfairness ($p = 0.026$, $n = 126$; two-sample median test; figure 2).

If the function of punishment is to deter harmdoers from imposing costs on oneself (i.e. to bargain for better treatment for oneself [10,13,20]) or others in the future—or even if its function is to enforce adherence to social norms [6]—then the punishment must be strong enough to erase unfairly gained benefits [1,35]. Otherwise, the harmdoer retains a net profit from the transgression, and thus will retain an incentive to continue to behave unfairly towards others in the future. Because unfair dictators took \$4 from recipients in round 1 of experiment 1, \$4 was also the minimum amount of punishment that would be expected to deter unfair dictators from behaving unfairly in the future. Third-party punishment of this magnitude was extremely rare: only 2 of 65 (3%) witnesses imposed at least \$4 worth of punishment on unfair dictators, a proportion no different from the proportion for witnesses of fairness (0 of 80; $p = 0.199$, Fisher's exact test). By contrast, 13 of 61 (21%) victims of unfairness punished at least \$4, a proportion significantly greater than that for both recipients of fairness (0 of 64; $p < 0.001$) and witnesses of unfairness ($p = 0.002$). Indeed, most victims of unfairness who punished (13 of 21) imposed at least \$4 worth of punishment.

According to the self-report measures of emotion, third parties did not become angry at unfairness: when controlling for envy (which was highly correlated with anger; $r = 0.637$, $p < 0.001$; see the electronic supplementary material, S1.5), a 2 (target: self, other) × 2 (treatment: fair, unfair) ANCOVA revealed a significant target × treatment interaction for anger ($F_{1,265} = 17.11$, $p < 0.001$). Witnesses of unfairness ($M = 0.533$, s.e. = 0.095, $n = 65$) did not report more anger than did witnesses of fairness ($M = 0.414$, s.e. = 0.084, $n = 80$; $p = 0.363$, partial $\eta^2 = 0.00$), but *victims* of unfairness ($M = 1.28$, s.e. = 0.098, $n = 61$) did report more anger than their fairly treated counterparts ($M = 0.418$, s.e. = 0.093, $n = 64$; $p < 0.001$, partial $\eta^2 = 0.13$; figure 3; see the electronic supplementary material, S1.3 and figure S1 for a replication with an implicit measure based on reaction time data). Thus, people became angry when treated unfairly but not when they only witnessed the unfair treatment of a stranger (cf. [36]). Importantly, this difference in the anger of witnesses versus victims of unfairness was not due to different perceptions of the transgression's fairness or moral wrongness (see the electronic supplementary material, S1.4 and figure S2).

Eleven of 65 witnesses of unfairness paid some cost to impose costs on the unfair dictator; with a much larger sample, one might argue, we therefore might have found statistical evidence for mild third-party punishment. However, because the witnesses of unfairness were not angry at the dictator (see above), we suspected that the predominant emotional response among witnesses of unfairness was envy, given that they had observed the unfair dictator obtain a higher payoff (\$9) than they themselves had received (\$5; see [37]). We found a significant target × treatment interaction for envy (with anger partialled out; $F_{1,265} = 4.53$, $p = 0.034$) witnesses of unfairness were more envious of the dictator than were the witnesses of fairness ($p \leq 0.001$, partial $\eta^2 = 0.07$). By contrast, victims of unfairness were no more envious than were their fairly treated counterparts ($p = 0.306$, partial $\eta^2 = 0.00$). Thus, had we observed a significant amount of third-party punishment among witnesses of unfairness, it plausibly could have been motivated by envy towards the unfair dictator rather than by moralistic anger. This difference in the emotions of the witnesses and victims of the dictator's unfairness explains why third-party punishment was quite rare and mild: witnesses of unfairness were envious of the dictator's ill-gotten gains—but not angry—and so they were likely to be reluctant to spend their own money to punish the dictator.

During experiment 1, we also ran a small fifth condition ($n = 45$; see the electronic supplementary material, S1.1) in which witnesses of unfairness started round 1 with \$9, enabling us to test whether witnesses' economic disadvantage relative to unfair dictators explained their surplus envy. Witnesses who started round 1 with \$9 were significantly less envious of unfair dictators (controlling for anger) than

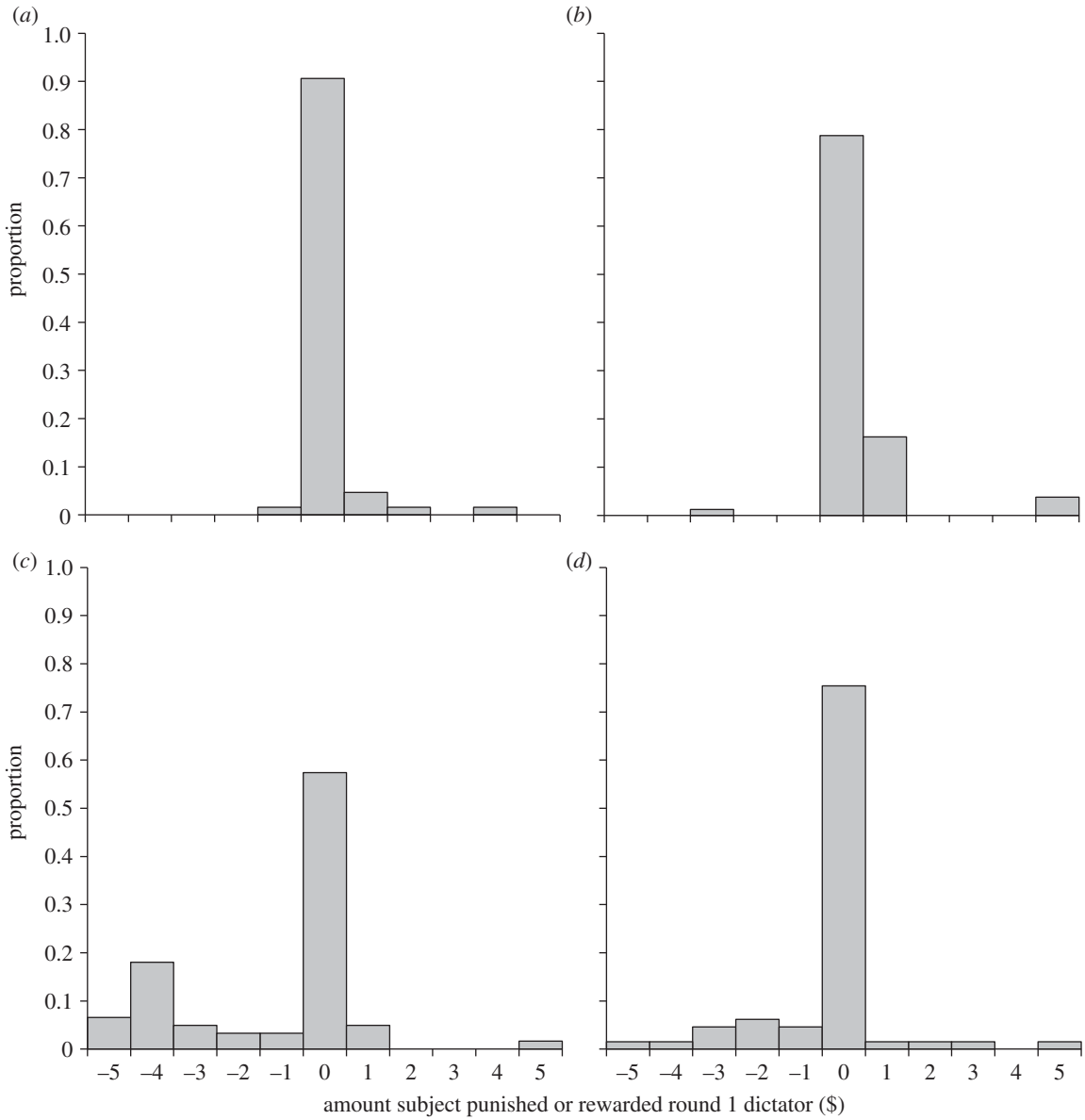

Figure 2. Punishment/reward distributions (experiment 1). Amount of money (in $) the subject punished (negative values) or rewarded (positive values) the round 1 dictator ($n = 270$). Values are in terms of the effect on the round 1 dictator, not the cost to the subject (cost-to-punish ratio = 1 : 4; cost-to-reward ratio = 1 : 1). Conditions: (*a*) fair–recipient, (*b*) fair–witness, (*c*) unfair–recipient and (*d*) unfair–witness.

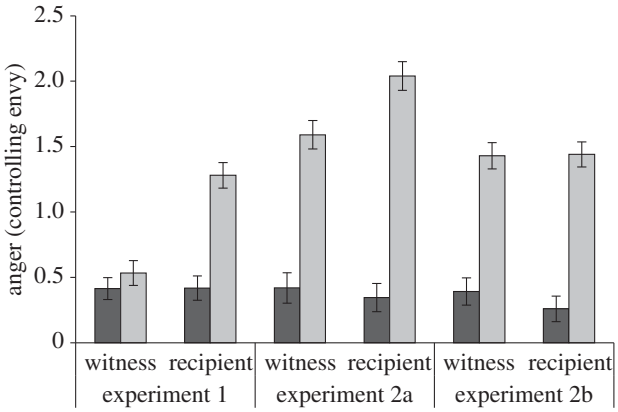

Figure 3. Self-reported anger (experiments 1, 2a and 2b). Self-reported anger (scale 0–5) towards dictator following round 1, controlling for envy ($n = 943$). Error bars = ±1 s.e. Light grey bars, unfair; dark grey bars, fair.

were witnesses who started with \$5 ($F_{1,107} = 8.35$, $p = 0.005$, partial $\eta^2 = 0.07$). Self-reported anger (controlling for envy) did not differ between groups ($F_{1,107} = 0.581$, $p = 0.448$, partial $\eta^2 = 0.01$). Witnesses of unfairness with \$9 did not punish an amount significantly different from zero ($z = -0.879$, $p = 0.379$, $n = 45$), nor differently than did the witnesses of unfairness with \$5 ($p = 0.475$, $n = 110$), even though doing so would have cost a smaller proportion of their stake. Thus, the emotional reactions of witnesses of unfairness were characterized by envy rather than moralistic anger.

(b) Experiment 2a

In experiments 2a (online sample) and 2b (undergraduate sample), we investigated how subjects' affective and behavioural forecasts in a hypothetical scenario would compare with the results from experiment 1 (see the electronic supplementary material, tables S1 and S2 for descriptive statistics for both experiments). This experiment was conducted not because we thought that participants' hypothetical responses would provide a reliable assay of how they would behave in a real-life situation, such as the one we explored in experiment 1, but because we wished to compare participants' *forecasts* of how they might behave and feel with participants' actual behaviour

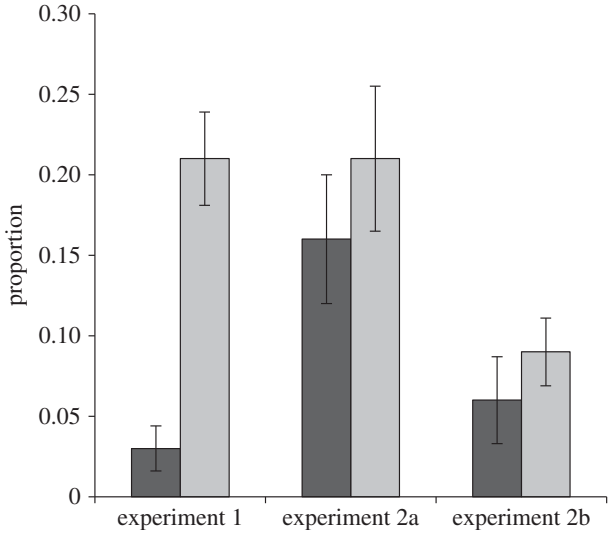

Figure 4. Punishment $\geq$ \$4 (experiments 1, 2a and 2b). Proportion of subjects in unfair conditions that punished (experiment 1) or reported they would punish (experiments 2a and 2b) $\geq$ \$4 ($n = 597$). Error bars $= \pm 1$ s.e. Light grey bars, recipient; dark grey bars, witness.

and emotional reactions in experiment 1. The rounds of the game were identical to experiment 1, except that subjects were instructed to report how they believed they would act and feel *in response to a hypothetical vignette*.

In experiment 2a—and in contrast to experiment 1—witnesses of hypothetical unfairness forecast that they would administer a greater-than-zero amount of punishment ($z = -2.38$, $p = 0.017$, $n = 137$), as did victims of hypothetical unfairness ($z = -2.66$, $p = 0.008$, $n = 148$; witnesses and victims of hypothetical unfairness did not differ significantly, $p = 0.391$). Four additional results suggest that a different psychological process was at work in experiment 2a than in experiment 1. First, witnesses of hypothetical unfairness forecast a much higher likelihood of punishing at least \$4 (the critical threshold for efficacious punishment) than did witnesses of hypothetical fairness ($p = 0.002$, Fisher's exact test). This proportion (22 of 137; 16%) was significantly larger than what we saw in the actual behaviour of experiment 1 subjects (2 of 65; 3%; $p = 0.009$, Fisher's exact test; figure 4). Second, the proportion of witnesses (16%) and victims (31 of 148; 21%) of hypothetical unfairness that forecast punishing at least \$4 did not differ ($p = 0.361$), contrary to experiment 1. Third, witnesses of hypothetical unfairness forecast a significant amount of anger towards unfair dictators in experiment 2a, again in contrast to experiment 1: there was a significant target $\times$ treatment interaction (controlling for envy; $F_{1,285} = 5.87$, $p = 0.016$). Witnesses of hypothetical unfairness ($M = 1.59$, s.e. $= 0.109$, $n = 133$) forecast more anger than did witnesses of hypothetical fairness ($M = 0.419$, s.e. $= 0.116$, $n = 147$; $p < 0.001$, partial $\eta^2 = 0.16$). Likewise, victims of hypothetical unfairness ($M = 2.04$, s.e. $= 0.110$, $n = 139$) forecast more anger than did recipients of hypothetical fairness ($M = 0.345$, s.e. $= 0.108$, $n = 141$; $p < 0.001$, partial $\eta^2 = 0.28$). Fourth, both witnesses ($F_{1,131} = 25.48$, $p < 0.001$, partial $\eta^2 = 0.16$) and victims ($F_{138} = 10.69$, $p = 0.001$, partial $\eta^2 = 0.07$) of hypothetical unfairness forecast significantly more anger (controlling for envy) towards the unfair dictator than their counterparts reported in study 1 (figure 3). These results are therefore consistent with proposals that norm violations elicit 'negative emotions' [4,6], which in turn motivate altruistic punishment, but here they resulted from affective forecasting rather than from responding to real-time events. (Recall, in contrast, that in experiment 1, which involved real-time behaviour and emotional responses rather than forecasting, no such moral outrage was found.)

(c) Experiment 2b

The pattern of punishment results for experiment 2b was virtually identical to those of experiment 2a: the proportion of witnesses of hypothetical unfairness that forecast they would punish at least \$4 (5 of 85; 6%) did not differ from that of victims of hypothetical unfairness (9 of 101; 9%; $p = 0.580$, Fisher's exact test)—a pattern that was similar to experiment 2a, but contrary to experiment 1. Interestingly, neither witnesses ($z = 0.183$, $p = 0.855$, $n = 85$) nor victims of hypothetical unfairness ($z = -0.162$, $p = 0.872$, $n = 101$) reported they would administer a greater-than-zero amount of punishment. Nevertheless, both witnesses of hypothetical unfairness ($p = 0.031$) and victims of hypothetical unfairness ($p = 0.012$) forecast they would punish significantly more than did their (hypothetically) fairly treated counterparts: this is because both witnesses ($z = 2.33$, $p = 0.020$, $n = 97$) and recipients of hypothetical fairness ($z = 2.98$, $p = 0.003$, $n = 85$) *rewarded* a greater-than-zero amount. Furthermore—and importantly—witnesses and victims of hypothetical unfairness did not forecast different amounts of punishment ($p = 0.743$). Thus, notwithstanding the fact that hypothetical punishment of unfairness appeared largely to have taken the form of withdrawing reward (rather than imposing costs) for subjects in experiment 2b, the punishment results largely replicated those obtained in experiment 2a (figure 4).

Moreover, the pattern of emotion-related results of experiment 2b was identical to that of experiment 2a: witnesses of hypothetical unfairness ($M = 1.43$, s.e. $= 0.100$, $n = 94$) forecast more anger (controlling for envy) than did witnesses of hypothetical fairness ($M = 0.392$, s.e. $= 0.104$, $n = 92$; $p < 0.001$, partial $\eta^2 = 0.13$). Likewise, victims of hypothetical unfairness ($M = 1.44$, s.e. $= 0.096$, $n = 109$) forecast more anger than did recipients of hypothetical fairness ($M = 0.259$, s.e. $= 0.098$, $n = 99$; $p < 0.001$, partial $\eta^2 = 0.16$). Witnesses ($F_{1,144} = 18.53$, $p < 0.001$, partial $\eta^2 = 0.11$) but not victims ($F_{157} = 0.005$, $p = 0.945$, partial $\eta^2 = 0.00$) of hypothetical unfairness also forecast significantly more anger (controlling for envy) towards the unfair dictator than the subjects in experiment 1 actually experienced (figure 3). The overall pattern of forecast behaviour and emotion in experiment 2b suggests that the students who were the subjects in experiment 2b had a slight tendency to believe that they would reward fair distributions, which the non-student subjects in experiment 2a did not share, but in every other way the results are identical to those of experiment 2a: subjects forecast that both experiencing and witnessing unfairness would cause them to become angry and to punish dictators to a greater extent than did subjects who forecast their responses to either receiving or witnessing fair treatment. Furthermore, both experiencers and witnesses of unfairness forecast equivalent likelihoods of punishing at least \$4.

4. Discussion

Experiment 1 indicates that, under the conditions we investigated, humans do not impose meaningful amounts of

third-party punishment on behalf of absolute strangers. The nominal and statistically non-significant amount of punishment we did observe was apparently motivated by envy because of a comparatively unfavourable personal outcome rather than by moralistic anger on behalf of a mistreated stranger. Our finding that the emotional reaction to witnessing unfairness is characterized by envy rather than moralistic anger is particularly inconvenient for the altruistic punishment hypothesis: to categorize a behaviour as an adaptation for altruistic benefit delivery, one needs to provide evidence that the psychological mechanisms that produce the behaviour in question have been designed for that specific function. That is, one needs to demonstrate that the behaviour is not caused by mechanisms designed for a different function. The presence of envy, rather than moralistic anger, in response to witnessing unfairness suggests that the psychological mechanisms involved in third-party punishment are, at least in part, designed to process cues that another individual has obtained better outcomes than oneself [38]. By contrast, we found no evidence that they are designed to process cues that an anonymous stranger has been harmed. We do not mean to imply that humans do not impose *any* third-party punishment: under some circumstances, they do [22,35,39]. However, our results cast doubt on the proposal that the mechanisms that motivate third-party punishment are *altruistic benefit-delivery systems* that are motivated proximately by moralistic anger.

Experiments 2a and 2b show furthermore that people inaccurately forecast their affective and behavioural responses to unfairness in experimental games: in particular, subjects who imagined themselves witnessing (rather than experiencing) unfair treatment forecast both more anger and punishment (and, in the case of experiment 2b, withdrawal of rewarding) than is observed among people who witness unfair treatment in the laboratory. This dissociation between hypothetical and actual third-party punishment raises the possibility that punishment imposed by mere witnesses of unfairness found in prior work resulted from demand characteristics, affective forecasting errors and the other methodological shortcomings we have cited here [8,25,27].

(a) Limitations

As mentioned at the outset, the goal of the experiments presented herein was not to systematically identify which specific methodological conventions were responsible for previous findings of third-party punishment on behalf of strangers. Rather, the goal was to test whether a suite of methodological conventions that are commonly applied within the third-party punishment game collude to create more third-party punishment in that experimental realization than would actually obtain in experiments that remediated those methodological shortcomings. As such, our results cannot determine with certainty the effects of particular aspects of previous designs, such as the strategy method. A recent survey of studies comparing the strategy method with the direct-response method across a variety of paradigms found that evidence surrounding the effect of using the strategy method is mixed [40], but, importantly, no study has been conducted to directly compare the amounts of third-party punishment elicited in experiments using the strategy method versus the direct-response method [40]. Therefore, further work is needed to determine the unique contribution of the use of the strategy method (and the other potential methodological artefacts we have identified herein) to the apparently exaggerated evidence for altruistic third-party punishment that previous work has revealed: we emphasize again that doing so was not our goal here. Despite this limitation, our results do strongly suggest that subjects' forecasts of their likely anger and punishment in response to witnessing unfairness in the standard third-party punishment game [6] are exaggerated.

(b) Conclusion

These findings are of broad significance in the study of human cooperation, because many researchers have proceeded under the assumption that altruistic punishment is a robust phenomenon that requires an adaptationist explanation. Indeed, two scientific 'problems' for which cooperation researchers over recent decades have been seeking adaptationist solutions might not be problems at all. Consider the puzzle framed by proponents of 'strong reciprocity,' such as Gintis [41], who claimed that humans are 'strong reciprocators' who are 'predisposed to cooperate with others and punish non-cooperators, even when this behaviour cannot be justified in terms of self-interest, extended kinship, or reciprocal altruism' (p. 169).

With respect to the former problem—'unjustified' predispositions to cooperate without apparent individual benefit—results from recent models suggest that a bias to cooperate, even when one faces cues of an interaction being one-shot, should be expected to coevolve with reciprocity. This is because mistaking a one-shot interaction for a repeated interaction is a less costly error than the reverse [42]. In terms of the latter problem (i.e. the claim that people punish non-cooperators even when the punisher does not stand to benefit individually), the results from experiment 1 call into question the claim that people engage in altruistic third-party punishment at all (see also [10,13]). We think another way forward in the study of third-party punishment in humans, as in the study of the evolved mechanisms that motivate human cooperation in general, is to intensify the search for direct or indirect benefits for punishers that outweigh the costs of punishment, consistent with all known cases of third-party intervention in non-human animals [43,44].

We thank Max Burton-Chellow, Steven Pinker, Stuart A. West and Richard W. Wrangham for their feedback on a previous draft, and David G. Rand for kind assistance with his data. Research supported by grants from the Air Force Office of Scientific Research (award no. FA9550-12-1-0179) to M.E.M and R.K., the Arsht Research on Ethics and Community Programme at the University of Miami to M.E.M. and an NSF Graduate Research Fellowship to E.J.P.

References

1. Clutton-Brock TH, Parker GA. 1995 Punishment in animal societies. *Nature* **373**, 209–216. (doi:10.1038/373209a0)
2. West SA, Griffin AS, Gardner A. 2007 Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *J. Evol. Biol.* **20**, 415–432. (doi:10.1111/j.1420-9101.2006.01258.x)
3. Bshary A, Bshary R. 2010 Self-serving punishment of a common enemy creates a public good in reef fishes. *Curr. Biol.* **20**, 2032–2035. (doi:10.1016/j.cub.2010.10.027)
4. Fehr E, Gächter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)

5. 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)
6. 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)
7. Henrich J *et al*. 2006 Costly punishment across human societies. *Science* **312**, 1767–1770. (doi:10.1126/science.1127333)
8. West SA, El Mouden C, Gardner A. 2011 Sixteen common misconceptions about the evolution of cooperation in humans. *Evol. Hum. Behav.* **32**, 231–262. (doi:10.1016/j.evolhumbehav.2010.08.001)
9. Burnham TC, Johnson DDP. 2005 The biological and evolutionary logic of human cooperation. *Anal. Kritik* **27**, 113–135.
10. McCullough ME, Kurzban R, Tabak BA. 2013 Cognitive systems for revenge and forgiveness. *Behav. Brain Sci.* **36**, 1–15. (doi:10.1017/S0140525X11002160)
11. Hagen EH, Hammerstein PR. 2006 Game theory and human evolution: a critique of some recent interpretations of experimental games. *Theor. Popul. Biol.* **69**, 339–348. (doi:10.1016/j.tpb.2005.09.005)
12. Kümmerli R, Burton-Chellew MN, Ross-Gillespie A, West SA. 2010 Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games. *Proc. Natl Acad. Sci. USA* **107**, 10 125–10 130. (doi:10.1073/pnas.1000829107)
13. Krasnow MM, Cosmides L, Pedersen EJ, Tooby J. 2012 What are punishment and reputation for? *PLoS ONE* **7**, e45662. (doi:10.1371/journal.pone.0045662)
14. McCullough ME, Pedersen EJ, Schroder JM, Tabak BA, Carver CS. 2012 Harsh childhood environmental characteristics predict exploitation and retaliation in humans. *Proc. R. Soc. B* 2**80**, 20122104. (doi:10.1098/rspb.2012.2104)
15. Carpenter JP, Matthews PH. 2013 Norm enforcement: anger, indignation, or reciprocity? *J. Eur. Econ. Assoc.* **10**, 555–572. (doi:10.1111/j.1542-4774.2011.01059.x)
16. Orne M. 1962 On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *Am. Psychol.* **17**, 776–783. (doi:10.1037/h0043424)
17. Dana J, Weber RA, Kuang J. 2007 Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Econ. Theor.* **33**, 67–80. (doi:10.1007/s00199-006-0153-z)
18. Nelissen RMA. 2008 The price you pay: cost-dependent reputation effects of altruistic punishment. *Evol. Hum. Behav.* **29**, 242–248. (doi:10.1016/j.evolhumbehav.2008.01.001)
19. Johnstone RA, Bshary R. 2004 The evolution of spiteful behavior. *Proc. R. Soc. Lond. B* **271**, 1917–1922. (doi:10.1098/rspb.2003.2581)
20. Sell A, Tooby J, Cosmides L. 2009 Formidability and the logic of human anger. *Proc. Natl Acad. Sci. USA* **106**, 15 073–15 078. (doi:10.1073/pnas.0904312106)
21. Lieberman D, Linke L. 2007 The effect of social category on third party punishment. *Evol. Psychol.* **5**, 289–305.
22. 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)
23. Bolton GE, Zwick R. 1995 Anonymity versus punishment in ultimatum bargaining. *Games Econ. Behav.* **10**, 95–121. (doi:10.1006/game.1995.1026)
24. Selten R. 1967 Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In *Beiträge zur experimentellen Wirtschaftsforschung* (ed. H Sauermann), pp. 136–168. Tübingen, Germany: JCB Mohr (Paul Siebeck).
25. Weber SJ, Cook TD. 1972 Subject effects in laboratory research: an examination of subject roles, demand characteristics, and valid inference. *Psychol. Bull.* **77**, 273–295. (doi:10.1037/h0032351)
26. Almenberg J, Dreber A, Apicella CL, Rand DG. 2011 Third party reward and punishment: group size, efficiency and public goods. In *Psychology of punishment* (eds NM Palmetti, JP Russo), pp. 73–92. Hauppauge, NY: Nova Science.
27. Wilson TD, Gilbert DT. 2005 Affective forecasting: knowing what to want. *Curr. Dir. Psychol. Sci.* **14**, 131–134. (doi:10.1111/j.0963-7214.2005.00355.x)
28. Cook KS, Yamagishi T. 2008 A defense of deception on scientific grounds. *Soc. Psychol. Q* **71**, 215–221. (doi:10.1177/019027250807100303)
29. Kawakami K, Dunn E, Karmali F, Dovidio J. 2009 Mispredicting affective and behavioral responses to racism. *Science* **323**, 276–278. (doi:10.1126/science.1164951)
30. Hareli S, Weiner B. 2002 Dislike and envy as antecedents of pleasure at another's misfortune. *Motiv. Emot.* **26**, 257–277. (doi:10.1023/A:1022818803399)
31. Reuben E, van Winden F. 2008 Social ties and coordination on negative reciprocity: the role of affect. *J. Public Econ.* **92**, 34–53. (doi:10.1016/j.jpubeco.2007.04.012)
32. Aronson E, Ellsworth P, Carlsmith J, Gonzales M. 1990 *Methods of research in social psychology*. New York, NY: McGraw-Hill.
33. Camerer C. 2003 *Behavioral game theory*. New York, NY: Princeton University Press.
34. Ernest-Jones M, Nettle D, Bateson M. 2011 Effects of eye images on everyday cooperative behavior: a field experiment. *Evol. Hum. Behav.* **32**, 172–178. (doi:10.1016/j.evolhumbehav.2010.10.006)
35. Petersen MB, Sell A, Tooby J, Cosmides L. 2010 Evolutionary psychology and criminal justice: a recalibrational theory of punishment and reconciliation. In *Human morality and sociality: evolutionary and comparative perspectives* (ed. H Høgh-Oleson), pp. 72–131. New York, NY: Palgrave Macmillan.
36. Batson CD *et al*. 2007 Anger at unfairness: is it moral outrage? *Eur. J. Soc. Psychol.* **1285**, 1272–1285. (doi:10.1002/ejsp)
37. Zizzo D, Oswald A. 2001 Are people willing to pay to reduce others' incomes? *Ann. Econ. Stat.* **63–64**, 39–65.
38. Price ME, Cosmides L, Tooby J. 2002 Punitive sentiment as an anti-free rider psychological device. *Evol. Hum. Behav.* **23**, 203–231. (doi:10.1016/S1090-5138(01)00093-9)
39. Phillips S, Cooney M, Carr T, Frady B. 2005 Aiding peace, abetting violence: third parties and the management of conflict. *Am. Sociol. Rev.* **70**, 334–354. (doi:10.1177/000312240507000207)
40. Brandts J, Charness G. 2011 The strategy versus the direct-response method: a first survey of experimental comparisons. *Exp. Econ.* **14**, 375–398. (doi:10.1007/s10683-011-9272-x)
41. Gintis H. 2000 Strong reciprocity and human sociality. *J. Theor. Biol.* **206**, 169–179. (doi:10.1006/jtbi.2000.2111)
42. Delton AW, Krasnow MM, Cosmides L, Tooby J. 2011 Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. **108**, 13 335–13 340. *Proc. Natl Acad. Sci. USA* (doi:10.1073/pnas.1102131108)
43. Raihani NJ, Grutter AS, Bshary R. 2010 Punishers benefit from third-party punishment in fish. *Science* **327**, 171. (doi:10.1126/science.1183068)
44. Smith JE, Van Horn RC, Powning KS, Cole AR, Graham KE, Memenis SK, Holekamp KE. 2010 Evolutionary forces favoring intragroup coalitions among spotted hyenas and other animals. *Behav. Ecol.* **21**, 284–303. (doi:10.1093/beheco/arp181)