Research Article

Psychological Science
2016, Vol. 27(3) 405–418

DOI: 10.1177/0956797615624469
pss.sagepub.com

Looking Under the Hood of Third-Party Punishment Reveals Design for Personal Benefit

Max M. Krasnow[1], Andrew W. Delton[2,3,4], Leda Cosmides[5,6], and John Tooby[5,7]

[1]Department of Psychology, Harvard University; [2]Department of Political Science, Stony Brook University; [3]College of Business, Stony Brook University; [4]Center for Behavioral Political Economy, Stony Brook University; [5]Center for Evolutionary Psychology, University of California, Santa Barbara; [6]Department of Psychology, University of California, Santa Barbara; and [7]Department of Anthropology, University of California, Santa Barbara

Abstract

Third-party intervention, such as when a crowd stops a mugger, is common. Yet it seems irrational because it has real costs but may provide no personal benefits. In a laboratory analogue, the third-party-punishment game, third parties ("punishers") will often spend real money to anonymously punish bad behavior directed at other people. A common explanation is that third-party punishment exists to maintain a cooperative society. We tested a different explanation: Third-party punishment results from a deterrence psychology for defending personal interests. Because humans evolved in small-scale, face-to-face social worlds, the mind infers that mistreatment of a third party predicts later mistreatment of oneself. We showed that when punishers do not have information about how they personally will be treated, they infer that mistreatment of other people predicts mistreatment of themselves, and these inferences predict punishment. But when information about personal mistreatment is available, it drives punishment. This suggests that humans' punitive psychology evolved to defend personal interests.

Keywords

cooperation, punishment, deterrence, evolutionary psychology, open data

Received 5/11/15; Revision accepted 12/7/15

A Crown Heights, New York, headline reports a curious incident: "Vigilant Bystanders Deliver Mugger to Hands of Justice" (2014). Why would third parties put themselves at risk to intervene in a conflict between strangers? In a city of anonymous millions, it is unlikely the third parties had any material stake in this conflict—their intervention appears irrational. Nonetheless, apparently irrational third-party intervention has been widely documented in laboratory and field settings (Buckholtz et al., 2008; Marlowe et al., 2011; McAuliffe, Jordan, & Warneken, 2015).

Here, we used an ecological-rationality approach to derive one explanation for why humans engage in third-party intervention: *the deterrence hypothesis*. A psychological mechanism is ecologically rational if it produces behavior that would, on average, have been adaptive in the environments that selected for its design. But that same mechanism may produce behavior that appears unwarranted or irrational when placed in evolutionarily atypical situations (Haselton et al., 2009; Kenrick et al., 2009; Lieberman, Tooby, & Cosmides, 2007; Mishra, 2014; Todd & Gigerenzer, 2007). According to the deterrence hypothesis, third-party intervention arises from a psychology designed to deter personally relevant mistreatment. This deterrence psychology exploits an enduring

Corresponding Authors:

Max M. Krasnow, Harvard University, Department of Psychology, 33 Kirkland St., Cambridge, MA 02138
E-mail: krasnow@fas.harvard.edu

Andrew W. Delton, Stony Brook University, Department of Political Science, Social and Behavioral Sciences Building, Stony Brook, NY 11794-4392
E-mail: andrew.delton@stonybrook.edu

regularity of ancestral social life: In small-scale social worlds, people who mistreat someone else now might also mistreat you (or others you value) in the future. This hypothesis contrasts with the *social-benefits hypothesis*, which proposes that third-party intervention exists to keep society running smoothly, regardless of whether such intervention costs the individuals who do it. According to this latter view, third-party intervention is targeted at people breaking society's rules, regardless of the implications for personal mistreatment.

Ecological Rationality and Human Third-Party Intervention

Human third-party intervention has been studied using the third-party-punishment game. Typically, one player is a dictator and is given a fixed stake of money (e.g., $10). The dictator can then choose to allocate some, none, or all of the stake to a recipient; whatever is not allocated, the dictator keeps. A third party, the punisher, learns about this decision and can spend money to reduce the dictator's earnings. Imagine that people are motivated to maximize their immediate economic gain. If the game is played once and anonymously, then rational third parties will never punish dictators, because they have no monetary incentive to do so. Dictators, rationally anticipating no punishment, should in turn keep their entire endowment. In contrast to these predictions, results have shown that punishers are often willing to punish dictators who give little to recipients; this has been found in the laboratory with industrialized populations, in the field with small-scale societies, and for both adults and children (Fehr, Fischbacher, & Gächter, 2002; Henrich et al., 2010; McAuliffe et al., 2015).

Why do human third parties appear to punish irrationally? The social-benefits hypothesis posits that third-party punishment is triggered when a fellow group member does wrong and that it functions to keep that group member providing benefits for the group—regardless of any personal costs for the punisher (e.g., Fehr & Fischbacher, 2004; Henrich et al., 2006, 2010). Despite its popularity, however, this hypothesis rests on assumptions about human evolution that may not be accurate (Krasnow, Cosmides, Pedersen, & Tooby, 2012; Krasnow & Delton, in press; West, Griffin, & Gardner, 2007). It also implies that the human mind is designed for the evolutionarily atypical societies people now inhabit, ones in which truly anonymous interactions are common. For instance, it assumes that punishment is designed to target people who are (a) in-group members and yet also (b) strangers you will never see again. While this often might be true in modern societies, it is unlikely to characterize human ancestral environments. And selection cannot create complex psychological adaptations, such as punishment, that are designed for conditions that did not long endure during human evolution (Tooby & Cosmides, 1992).

We suggest instead that this puzzle can be parsimoniously solved by considering the evolutionary history and natural ecology of human third-party punishment. The human mind is a collection of inherited neural decision-making mechanisms that were organized to make what were fitness-enhancing choices given the conditions—the social and informational ecology—that prevailed during their evolution (Tooby & Cosmides, 1992). These designs interpret current conditions and signals in terms of their ancestral, rather than their modern, significance. Such architectures often use informative cues, even though such cues are sometimes fallible (e.g., Delton, Krasnow, Cosmides, & Tooby, 2011; Krasnow, Delton, Tooby, & Cosmides, 2013; Petersen, Sell, Tooby, & Cosmides, 2012; Todd & Gigerenzer, 2007). For example, men can become aroused by women they know to be using contraception. Does this imply that sexual arousal did not evolve to promote reproduction? Obviously not: Ancestrally, visual appearance cues, not seeing a woman take a pill, carried reliable information about fertility—explaining modern but "irrational" arousal responses.

By the same token, although punishment is not logically warranted based on the evolutionarily atypical particulars of standard experiments—anonymous dictators, anonymous punishment, a one-shot interaction certain never to be repeated—it may nonetheless trigger a motivational strategy designed for more evolutionarily typical conditions. Humans evolved in relatively small social worlds numbering in the tens, or more rarely hundreds (Dunbar, 1993). Thus, a person you witness treating others poorly can later target you or your friends or family. Even a stranger just passing through is someone you might see again; the mere fact that you are meeting them now predicts seeing them again (Krasnow et al., 2013). The problem is especially acute if the poor treatment resulted from a stable feature of the malefactor (e.g., aggressive personality) rather than a transient or idiosyncratic event (e.g., a one-off argument). Punishing bad behavior directed toward others can deter poor treatment of you and those you value, yielding both direct and indirect benefits (Raihani, Grutter, & Bshary, 2010; Roos, Gelfand, Nau, & Carr, 2014). Moreover, refraining from punishing may be interpreted to mean that you fear or defer to the malefactor, which could lead that person and others to feel freer to exploit you in the future.

Is this long-standing correlation between malefactors' mistreatment of other people and the potential for them to mistreat you and people you value incorporated into the evolved design of the motivational mechanisms that generate third-party punishment? If it is, then people may punish as third parties even when such punishment is

not rationally warranted (e.g., in anonymous experiments or modern mass societies). This would follow because—in small-scale ancestral environments—a plausible heuristic design for protecting yourself and people you value is to be motivated to punitively recalibrate those who feel free to exploit others in your presence. If this theory is correct, third-party punishment may be a proximate reaction to third-party mistreatment, but it actually stems from a mechanism that was ultimately designed for first-person deterrence.

The Present Research

To contrast these hypotheses, we conducted two experiments using a modified third-party-punishment game. As usual, dictators divided their endowment between themselves and a recipient, and punishers decided how much to punish contingent on the dictator's division. The novel twists were that (a) we assessed, using a different task, how much the dictator valued the recipient and, separately, the punisher, and (b) we asked punishers to infer how much the dictator valued the recipient and the punisher. This design allowed us to test three predictions from the deterrence hypothesis. First, dictators' allocations to recipients would predict how much they value a third party, the punisher. Second, third-party punishers would correctly infer that dictators' treatment of recipients would predict how much dictators valued the punisher. Third, punishers' inferences about how they are personally valued by the dictator would predict their actual punishment.

In Study 2, we further modified the design to explore two situations that contrast predictions of the two hypotheses, comparing a condition in which punishers learn only how dictators treat recipients with a condition in which punishers learn how a dictator treated a recipient and the punishers themselves. According to the deterrence hypothesis, punishers' minds treat a dictator's behavior toward the recipient as a cue to how the dictator will treat the punisher in the future. This cue—treatment of an unknown third party—should have less weight in decision making if a better cue—treatment of the self—is available. This contrasts with the predictions of the social-benefits hypothesis. For Study 1, the social-benefits hypothesis predicted no special connection between third-party treatment and inferences about how the punisher would be treated, or between inferences about personal treatment and punishment. In Study 2, the difference was even starker: Whereas the deterrence hypothesis predicted that anonymous third-party information should be largely ignored when personal treatment is available, the social-benefits hypothesis predicted that information about third-party treatment and personal treatment should be interchangeable. The social-benefits account posits that what matters is that one group member is offending another group member, not whether the person witnessing the offense could be next.

Study 1

Method

Subjects. One hundred twenty people (72 women, 48 men) participated in this study. Our primary analyses were correlation tests and within-subjects contrasts, the latter conducted in a within-subjects analysis of variance (ANOVA). Given moderate effect sizes for both, a correlation test would be less powerful. Therefore, we focused on it for power considerations. For a two-sided alpha (α) of .05, a power of .85, and a rho (ρ) of .3, the necessary sample size was approximately 95 subjects (Faul, Erdfelder, Buchner, & Lang, 2009). Because we had more funds than needed to run 95 subjects, we increased our desired sample size to 120, which conveniently yielded 40 groups of 3 subjects each. Subjects were drawn from the University of California, Santa Barbara, psychology study pool and were primarily undergraduates along with a few community members. They received money on the basis of decisions made during the experiment, on top of a $5 show-up fee. Because of a program crash, data from 1 person were lost, which left 119 individuals for analysis. No deception was used at any point in this study, and the protocol was approved by the University of California, Santa Barbara, Institutional Review Board.

Procedure. Experimental sessions were conducted in a small laboratory with semiprivate cubicles. Each session had 6 or 9 same-sex subjects divided into groups of 3 each. Each group played an anonymous, one-shot third-party punishment game. One subject was assigned to be the dictator, one to be the recipient, and one to be the punisher. However, subjects did not know which role they were assigned to. Instead, each subject committed to the decisions they would make as punisher and then committed to the decisions they would make as dictator (see Game Play). Only after subjects committed to decisions for each role did they learn what role they were assigned to and what their payouts would be. This procedure allowed us to maximize the amount of data collected by gathering punishment data from every participant.

Third-party-punishment game: division and punishment. In the third-party-punishment game, the dictator in each group received $10 and decided how much of that money to keep and how much to give to the recipient. Dictators were constrained to divisions in increments of $2.50 (i.e., allowable divisions, or *levels of allocation*,

were \$0-\$10, \$2.50-\$7.50, \$5-\$5, \$7.50-\$2.50, and \$10-\$0, or 0%, 25%, 50%, 75%, and 100% allocation, respectively). Recipients had no say over this division and could only passively accept the amount they were allocated. Punishers received a separate sum of \$5 and could spend none of it, all of it, or part of it in \$1 increments to reduce the money, in 20% increments, the dictator kept (see Henrich et al., 2010; punishment decisions were made using the strategy method; see Game Play for details). For instance, if the dictator chose to keep \$7.50 (giving the recipient \$2.50) and the punisher spent \$2, then the dictator lost \$3 of that \$7.50 allocation (i.e., 40% of \$7.50). In this example, the dictator would then have \$4.50, the recipient \$2.50, and the punisher \$3. To avoid using directive language, we designed the materials for subjects so they referred to dictators and punishers as "allocators" and "responders," respectively.

Valuation and inferences about valuation. In their role as dictators, subjects completed a task that captured how much they valued the recipient and the punisher (see Delton & Robertson, 2016). For each of these two targets, dictators were asked to make 12 binary decisions between allocating a certain amount of money to themselves or a different amount of money to another person. Within each series of 12 decisions, the amount for the recipient or punisher was held constant, but the amount for dictators was randomly varied (see Table S1 in the Supplemental Material). For instance, a dictator might first be asked to decide between keeping \$10 or giving \$20 to the recipient, then keeping \$3 or giving \$20 to the recipient, then keeping \$12 or giving \$20 to the recipient, and so forth. Dictators made their choices knowing that the experimenter would randomly select and pay out only 1 of these 12 decisions for each target at the end of the session (in addition to their other earnings).

To assess punishers' inferences about how much dictators valued the recipient and themselves, we showed subjects in their role as punishers the same set of decisions that dictators were asked to make. Punishers selected the choices they believed the dictator had made, once with respect to the recipient and once with respect to themselves. Punishers earned \$0.25 for every correct answer.

Dictators' decisions were combined to create two valuation scores, one each for the recipient and the punisher (Delton, 2010; Kirkpatrick, Delton, Robertson, & de Wit, 2015). Similar valuation scores were computed for punishers' inferences. In brief, the valuation scores were created by looking for *switch points*. For example, if the dictator chose to keep \$12 instead of giving \$20 to the recipient, but gave up the opportunity to get \$10 to deliver \$20 to the recipient, then the valuation score was calculated as the average of the ratios bounding this switch point, (12/20 + 10/20)/2 = .55. Perfect consistency was not required; the actual scoring method computed the best-fitting switch point (see Delton, 2010). Given the decisions we used, valuation scores could range from .03 to .75, with greater numbers representing a greater valuation by the dictator of the other person. For instance, if a valuation score for a punisher's inference was .25, this meant that the punisher assumed that the dictator would forgo up to \$5 to give \$20 to the other person. If the valuation score was .50, this meant that the punisher assumed that the dictator would forgo up to \$10 to give \$20 to the other person.

Game play. Subjects learned about all aspects of the game (i.e., the third-party-punishment game and the valuations), and they answered comprehension questions before they could begin play. They knew that they were one member of a triad, but they did not know the identities of the other two members. Subjects made both dictator and punisher decisions without knowing which ultimate role they would be assigned when it came time to calculate their earnings.

Subjects were first asked to take the role of punisher. By having subjects make punisher decisions prior to dictator decisions, we prevented their punishment decisions from being contaminated by decisions they made in the dictator role. In addition, the experiment was conducted using the *strategy method*. That is, when making decisions as the punisher, subjects did not have access to the dictator's actual decisions. Instead, they considered in turn each possible division of money that the dictator might choose and committed to punishment responses given that possible division. In addition, for each possible dictator division, punishers also made inferences about the dictator's valuations of the recipient and the responder. (Only the punisher's punishment choice for the division the dictator actually made affected payouts at the end of the experiment; the remaining four punishment choices were analyzed as data but played no role in payouts.)

Punishers committed to punishment and made valuation inferences about each possible dictator division in either ascending or descending order. For example, in the ascending order, punishers were asked to start by assuming that the dictator allocated nothing to the recipient. Given this, how much did the dictator value the recipient and punisher? Punishers answered by completing the recipient- and punisher-valuation tasks as they believed the dictator would have—on the assumption that the dictator had allocated nothing to the recipient. Then punishers decided how much they would punish, again on the assumption that the dictator allocated nothing to the recipient. These punishers then repeated the valuation inferences and punishment task, now assuming the dictator had allocated \$2.50, \$5.00, \$7.50, and \$10.00, in turn.

Subjects were next asked to take on the role of dictator. They completed the valuation task with respect to both the recipient and the punisher. Then, still as dictators, they completed the dictator decision for the third-party-punishment game, dividing $10 between themselves and the recipient. Finally, they learned what role the experimenter had assigned them in their triad for the purpose of paying out their earnings (dictator, recipient, or punisher). They also learned what decisions the chosen dictator and punisher in that triad had made, which allocations would be actualized, and the amount of money they earned.

Since many of the decisions in strategy method games are never realized, there has been debate concerning how much the results of these games can compare with their traditional-form counterparts (in which, for example, dictators decide first and punishers are informed of their actual decision before deciding on punishment). While some deviation from traditional-form games has been observed (Brosig, Weimann, & Yang, 2003), in large part the use of the strategy method has not been found to be a serious experimental contaminant (Brandts & Charness, 2000). Consistent with this, a recent article on the third-party-punishment game shows that behavior in this game is not affected by whether normal gameplay or the strategy method is used (Jordan, McAuliffe, & Rand, 2015). We note also that the strategy method was used in some of the most high-impact work done from a social-benefits perspective, even with nonindustrialized populations (Henrich et al., 2006).

Using the strategy method allowed us to adequately measure punishers' inferences and punishment behavior for choices that dictators seldom made. While this entailed punishers making many decisions—five punishment decisions and 120 valuation inferences (12 inferences × 5 levels of allocation × 2 targets of trade-off: recipient/punisher), previous work with the valuation task has found that subjects can make such decisions intuitively in several seconds and that their decisions are made reliably across hundreds of questions (Delton, 2010). In other research on this and similar tasks, people made trade-offs similarly whether rewards were real or hypothetical (Delton, 2010; Locey, Jones, & Rachlin, 2011). Finally, choices on the valuation task have been shown to predict behavior in a variety of other contexts, both in the real world and in laboratory games (see Delton & Robertson, 2016, for a review).

Results

Descriptive statistics for punishment, allocations, and valuation. Behavior in our modified third-party-punishment game resembled behavior in typical versions. As usual, punishers were sometimes willing to punish (Fig. 1). For instance, when assuming that dictators allocated nothing to recipients, 55% of punishers (65 of 119) spent at least $1 on punishment, and 15% (18 of 119) spent their entire $5 on punishment. Dictators were sometimes willing to allocate money to recipients (diamonds in Fig. 1). For instance, 64% of dictators (76 of 119) were willing to allocate at least some money to

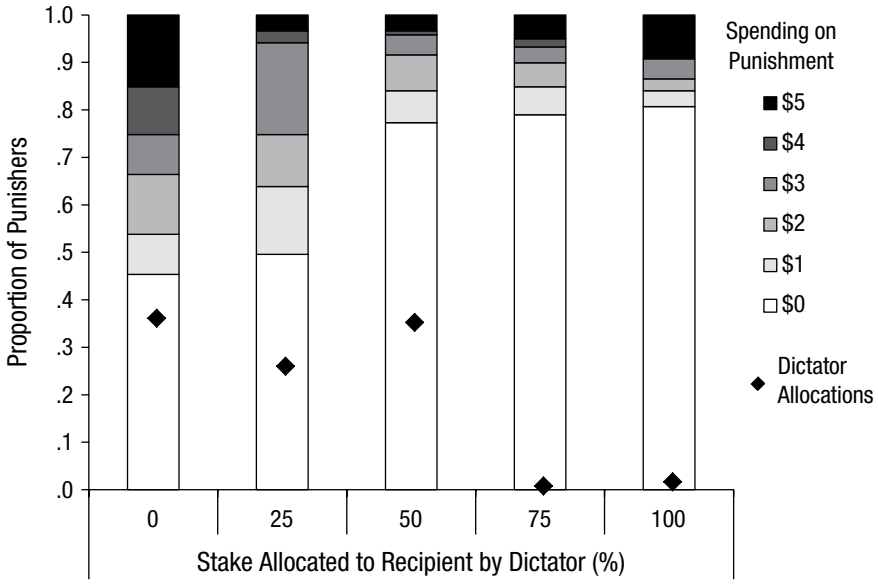

Fig. 1. Punisher spending and dictator allocation in Study 1. Bars indicate the proportion of punishers who spent $0 to $5 on punishing as a function of how much of the stake dictators allocated to recipients. Diamonds indicate the proportion of dictators who allocated that amount of the stake to the recipient.

recipients, and 38% (45 of 119) were willing to allocate at least half of the total stake. Even when dictators allocated half or more of their stake to recipients, a minority of punishers were still willing to punish. This might reflect noise, as subjects who did not anticipate such decisions being actualized may have answered with less concern, or it might reflect antisocial punishment, which has been observed in other games (Herrmann, Thöni, & Gächter, 2008; Masclet, Noussair, Tucker, & Villeval, 2003). When dictators allocated all $10, punishment, though still costly, was completely ineffective; in this case, dictators had no residual endowment that punishment could reduce. Nonetheless, past experiments have also shown that people are willing to pay for costly punishment even when such punishment is completely ineffective (Yamagishi et al., 2009).

Dictators moderately valued recipients and punishers (*M*s = .22 and .23, respectively). Roughly, a dictator would forgo receiving up to $4 if recipients and punishers could receive $20.

Dictator allocations predict behavior in other situations. A basic assumption of the deterrence hypothesis is that dictators' behavior in one situation (e.g., the division of money between themselves and the recipient) predicts their behavior in another situation (e.g., the valuation task). Our results confirmed this hypothesis: Dictators who allocated less to the recipient also valued their recipient less, $r = .40$, $p < .001$, 95% confidence interval (CI) = [.24, .54].[1] It would be surprising on many accounts if this were not the case.

Dictator allocations to recipients predict how much dictators value punishers. Although the results reported so far might be obvious, it is not obvious that how much dictators allocate to one person should predict how much they value a different person. Nonetheless, dictators who allocated less to the recipient valued the punisher less as well, $r = .43$, $p < .001$, 95% CI = [.27, .57]. Of equal significance, the more that dictators valued the recipients, the more they valued the punishers, $r = .84$, $p < .001$, 95% CI = [.78, .87]. Thus, allocation to recipients contained information that a punisher could use to infer how the dictator would treat them personally.

Punishers assume that dictators' treatment of recipients predicts how much dictators value punishers. As we have shown, dictators' divisions toward recipients predict how much dictators value punishers. But do third parties actually make inferences using this cue? This is a primary prediction of the deterrence hypothesis, and our results confirmed it: As shown in Figure 2a, the more that dictators allocated to recipients, the more they were thought to value both the punisher and the recipient. In fact, the inferences punishers made for the two targets (themselves and recipients) were indistinguishable. Using a 2 × 2 repeated measures ANOVA with allocation and target as factors, we found that the effect of allocation on inferences was significant, $F(2.32, 273.72) = 18.15$, $p < .001$, $\eta^2 = .13$, but there was no effect of target and no interaction between target and allocation (*p*s = .25 and .17; see Table S2 in the Supplemental Material).

These results are telling: After learning specific information about how the dictator would treat another person—the recipient—punishers drew no distinction between behavior they expected from dictators personally and behavior they expected dictators to direct toward recipients. At all five levels of allocation, punishers' inferences about dictators' valuation of punisher and recipients were highly correlated, *r*s = .80, .65, .76, .71, and .81 for 0%, 25%, 50%, 75%, and 100% allocation, respectively; *p*s < .001. We also note that punishers' inferences about dictators' trade-offs closely matched dictators' actual trade-offs (cf. the dashed and solid lines in Fig. 2a).

Inferred valuation predicts third-party punishment. Consistent with past research, our results showed that when dictators gave less to recipients, they received more third-party punishment (Fig. 2a). This was revealed by a repeated measures ANOVA, which showed that the amount of money allocated to recipients accounted for 17% of the variance in punishment decisions, $F(3.27, 385.45) = 24.57$, $p < .001$, $\eta^2 = .17$ (see Table S3 in the Supplemental Material).

But why? According to the deterrence hypothesis, third-party punishment reflects the punishers' inferences about how they themselves would be treated by dictators. This was the case: As shown by a hierarchical linear model with random slopes and intercepts, when punishers thought they were valued less by dictators, they punished dictators more, $\gamma = -1.10$, $t(118) = -2.47$, $p < .05$ (see Table S4 in the Supplemental Material).

Given that punishers' inferences about how dictators value punishers and recipients were indistinguishable, it is not surprising that a hierarchical linear model also showed that third-party punishment was predicted by punisher's inferences about dictators' valuation of recipients, $\gamma = -1.38$, $t(118) = -3.64$, $p < .001$ (see Table S5 in the Supplemental Material). (Considering potential measurement error and the high correlations between the two measures, averaging .75, these two measures are probably best conceptualized as capturing only a single psychological dimension in this study.)

More generally, all of our measures correlated strongly with one another. When dictators highly valued recipients, they also highly valued punishers ($r = .84$). When punishers' inferred that dictators valued recipients, punishers also inferred that dictators valued punishers (all *r*s > .65). And as shown in Figure 2a, punishers' inferences

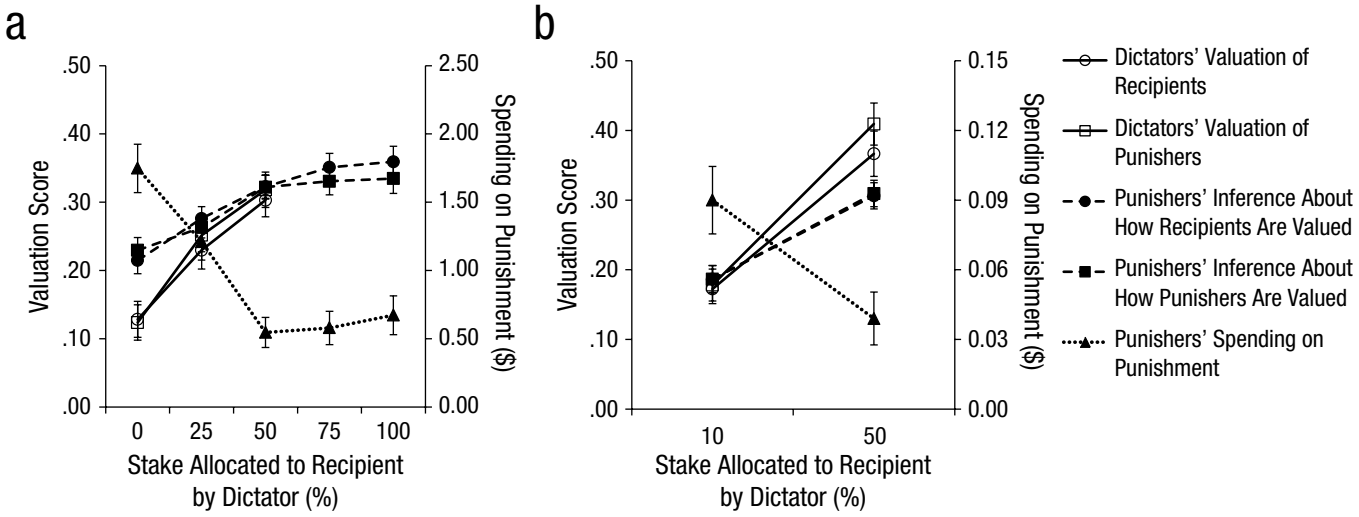

Fig. 2. Valuations and punisher spending in (a) Study 1 and (b) the matching (one-recipient) condition of Study 2. The *y*-axes on the left of each graph show dictators' mean valuations of recipients and punishers, along with punishers' mean inferences about dictators' valuations. The *y*-axes on the right of each graph show the mean amount punishers spent to punish dictators. All values are graphed as a function of the percentage of the total stake that dictators allocated to recipients. So few Study 1 dictators allocated more than 50% ($n = 3$) that we did not include their valuations in these graphs. In (b), the lines for punishers' inferences completely overlap. Error bars indicate ±1 *SEM*.

of dictators' valuations closely tracked dictators' actual valuations. Further, third-party punishment was almost a mirror image of punishers' inferences of how much dictators valued the punisher and recipient.

As predicted by the deterrence hypothesis, these findings showed that punishers assume that how dictators treat an anonymous third party will predict how dictators will treat the punisher, and that predicts how much dictators are punished. The social-benefits hypothesis makes no predictions about how the dictator values the third-party punisher, nor about the punisher's inferences about valuation, which leaves important features of these results unexplained.

Study 2

Method

Subjects. Four hundred people (188 women, 212 men) were recruited from Amazon's Mechanical Turk (MTurk) to participate in this study (average age = 34.85 years, *SD* = 10.96). Given the power analysis used in Study 1, we found that a sample size of 100 subjects per condition would provide adequate power to detect medium-sized within-condition effects, so we decided to recruit 400 subjects. This yielded a convenient 100 groups of 4 subjects each. Unlike in Study 1, subjects were instructed in only one role; they knew prior to making decisions whether they were a dictator or punisher. One hundred people participated as dictators, and 300 participated as punishers in one of three conditions (100 subjects per condition). Recruitment was limited to U.S. residents with a prior MTurk approval rating of 95% or higher. They received money on the basis of decisions made during the experiment on top of a $0.50 baseline fee. Because of technical difficulties, data from 5 people were unavailable, which left 395 individuals for analysis. No deception was used at any point in this study, and the protocol was approved by the Harvard University Institutional Review Board.

Procedure. Subjects participated in Study 2 in groups of four. Each group had one dictator and three punishers. Dictators were aware that all three other players were punishers. For each punisher, dictators were endowed with $1.00 and decided to allocate either $0.10 or $0.50 of this amount to the punisher (thus, all punishers were also recipients). Dictators also completed the valuation task in regard to each punisher. As in Study 1, valuation scores ranged from .03 to .75, but the stakes were lower, as appropriate to MTurk compensation (see Table S6 in the Supplemental Material for the choices). In every condition, punishers inferred how much the dictator valued two people: the punishers themselves and one other recipient of the dictator's allocation. In all conditions, punishers were endowed with a single sum of $0.50 and could spend this money in $0.10 increments to reduce the amount, in 20% increments, that the dictator kept. As in Study 1, punishers inferred dictators' valuations about themselves and one recipient; each correct guess earned $0.02. Dictators were aware that all three punishers could punish them and which of their own allocations punishers could condition that punishment on.

Each of the three punishers in a group was assigned to a different condition (conditions are detailed in the following subsections). These conditions allowed us to test whether information about treatment of oneself overrides information about treatment of another person. The self-and-other condition was our key experimental condition: In this condition, punishers learned how the dictator treated them and how the dictator treated another person. This allowed us to test whether treatment of the self was privileged. The two-recipients condition was our primary control condition: In this condition, punishers learned how the dictator treated two other people, but—importantly—punishers did not learn how they themselves were treated. This allowed us to test whether any pattern of results in the self-and-other condition was unique or occurred merely because the punisher learned about the dictator's treatment of two other people rather than one other person. The one-recipient condition served as a baseline and replication of Study 1: In this condition, the punisher learned only about how the dictator treated a single third party.

One-recipient condition. From the perspective of punishers in the one-recipient condition, the game was a standard third-party-punishment game. Punishers in this condition were aware only of three players: themselves, the dictator, and a "recipient." They did not know that this recipient was also a punisher (or that a third punisher existed). Importantly, punishers in this condition did not know that they themselves were the target of an allocation by the dictator (this money was described as a variable participation bonus). For both possible divisions between the dictator and the recipient—$0.50-$0.50 or $0.90-$0.10 favoring the dictator—punishers inferred how much the dictator valued the recipient and themselves and decided on a punishment for that allocation. As in Study 1, smaller allocations should lead these punishers to infer lower valuations by the dictator, and these inferences should produce more punishment.

Two-recipients condition. From the perspective of punishers in the two-recipients condition, the game was a third-party-punishment game, but with two recipients instead of one. Punishers were aware that all four players existed: themselves, the dictator, and two recipients (A and B). As in the one-recipient condition, punishers in this condition did not know that recipient A and recipient B were also punishers, or that they themselves were the target of an allocation by the dictator (as before, this money was described as a variable participation bonus). The two recipients were not treated identically: Punishers made valuation inferences with respect only to recipient A. Punishers made inferences and committed to punishment for all four combinations of dictator divisions: 50-50 for recipient A and dictator-favoring for recipient B, dictator-favoring for recipient A and 50-50 for recipient B, 50-50 for both divisions, and dictator-favoring for both divisions.

According to both the deterrence and social-benefits hypotheses, dictators' treatment of recipients A and B will regulate third-party punishment. But punishers should not weigh the dictator's treatment of recipient A over recipient B (or vice versa) when deciding how much to punish.

Three patterns would be consistent with equal weighting of treatment of both recipients. Punishment might (a) increase with every instance of poor treatment, (b) increase only if both recipients are treated poorly, or (c) increase as long as at least one recipient is treated poorly, with no additional increase if the other is treated poorly. The deterrence hypothesis predicts that the same pattern should hold for inferences of valuation.

Regardless of which pattern holds, the identity of the recipient (A or B) should not influence responses. Which recipient is treated poorly should matter a great deal, however, when one of the two recipients is also the punisher. This hypothesis was tested in the self-and-other condition.

Self-and-other condition. From the perspective of punishers in the self-and-other condition, the game was not a pure third-party-punishment game. Punishers in this condition were aware of three players: themselves, the dictator, and a recipient. As in the other two conditions, they were not aware that the recipient was also a punisher. The key difference between this condition and the others is that punishers here knew that they were also the target of an allocation by the dictator.

Whereas punishers in the two-recipients condition had information about how their dictator treated two third parties—two people other than themselves—punishers in the self-and-other condition had information about how their dictator treated a third party and how their dictator treated the punisher personally. Punishers made inferences and committed to punishment for all four combinations of dictator divisions: 50-50 for the punisher and 10-90 for the recipient (in favor of the dictator), 10-90 for the punisher (in favor of the dictator) and 50-50 for the recipient, 50-50 in both divisions, and 10-90 (in favor of the dictator) in both divisions.

The social-benefits hypothesis predicts that treatment of the other recipient will affect punishment even when punishers know how they themselves were treated. After all, according to this hypothesis, the function of third-party punishment is to punish bad behavior even when there are no personal costs or benefits at stake. The deterrence hypothesis predicts that treatment of the punisher will be the main factor in determining punishment:

Punishment should depend primarily on whether the punisher was treated poorly and only partly, if at all, on whether the recipient was treated poorly. According to the deterrence hypothesis, third-party punishment exists to defend the self and valued other people, not society as a whole. Although previous research shows that people punish more when they themselves are treated poorly, compared with merely observing poor treatment of a third party (Fehr & Fischbacher, 2004), the present study was the first to put these two cues in direct competition with each other and measure their effect on downstream inferences. It was therefore also the first study to test these discriminative predictions of the social-benefits and deterrence hypotheses.

Results

Descriptive statistics for punishment, allocations, and valuation. As in Study 1, behavior in our modified third-party-punishment games resembled behavior in typical versions of the game. Again, punishers were sometimes willing to punish dictators who allocated little money: For example, when dictators kept \$0.90 and gave \$0.10, they were punished approximately 35% of the time. And dictators were sometimes willing to allocate money: Dictators split their endowment evenly approximately 50% of the time. (See Fig. S1 in the Supplemental Material for complete descriptive data on allocations and punishments by condition.) Also as in Study 1, dictators moderately valued other subjects (M = .28). Roughly, a dictator would forgo up to \$0.28 if a punisher could receive \$1.

These basic results from Study 2 also suggest that using the strategy method did not dramatically affect our data: There were many fewer hypotheticals to consider in Study 2 than in Study 1, yet people's behavior was similar across both studies (despite changes in both sample and setting).

Dictators' treatment and valuation of one punisher predicts treatment and valuation of other punishers. As in Study 1, dictators' allocations to one punisher predicted how much they valued that punisher, rs > .45, ps < .001 (Table S7 in the Supplemental Material). Further, when dictators made allocations, the more money they gave to one punisher, the more they allocated to the other punishers (rs > .84, ps < .001). And the more that dictators valued one punisher, the more they valued the other punishers (rs > .88, ps < .001; Table S7). Unsurprisingly given these correlations, dictators' allocations to one punisher predicted how much they valued the other punishers (rs > .46); these between-punisher correlations were, in fact, just as strong as the within-punisher correlation (Pearson's zs < 0.6, ps > .55; Table S7).

When they have no other information, punishers punish on the basis of how third parties are treated. According to the deterrence hypothesis, when all recipients are third parties (as in the one-recipient and two-recipients conditions), low allocations by dictators will elicit more punishment from punishers because punishers will use this cue to infer that dictators do not value them highly. Punishers' inferences of how much dictators value them are based on their inferences of how much dictators value the other recipients (which are based on how much the dictator allocated to the recipients). The social-benefits hypothesis shares the punishment prediction but is silent regarding inferences about how highly the dictator values the punisher.

Both effects obtained (see Fig. 2b). When the dictator treated the recipient poorly, punishers in the one-recipient condition punished more (mean difference = \$0.051, SD = 0.136, 95% CI = [0.024, 0.078]), $t(99) = 3.751$, $p < .001$, $d = 0.37$. Punishers also inferred that they would be valued less by the dictator in such cases (mean difference = .123, SD = .205, 95% CI = [.082, .164]), $t(99) = -6.00$, $p < .001$, $d = 0.60$. Punishers' inferences about how they were valued were closely associated with inferences about how much the dictator valued the third party: Punishers in this condition assumed that dictators valued recipients less when the dictator treated the recipient poorly (mean difference = .120, SD = .203, 95% CI = [.079, .160]), $t(99) = 5.89$, $p < .001$, $d = 0.59$.

In the two-recipients condition, both hypotheses predicted that punishers would not privilege treatment of one recipient over another when deciding how much to spend punishing dictators. This could mean that punishment rises steadily with each additional act of poor treatment, rises only when both recipients receive poor treatment, or rises with the first act of poor treatment but not anymore with the second. The actual results revealed a combination of the first two patterns. As Figure 3a shows, the primary effect appears to be that both recipients must be treated poorly in order for punishers to increase spending. This should be reflected statistically in an interaction term, but the interaction was only marginally significant in a 2 × 2 repeated measures ANOVA (see Table S8 in the Supplemental Material), $F(1, 95) = 3.151$, $p = .079$, $\eta^2 = .03$. Even if the interaction is disregarded, the main effects show that punishment increases to roughly the same extent regardless of which recipient was treated poorly—there was a main effect of poor treatment both for recipient A, whom the punishers also made inferences about, $F(1, 95) = 6.013$, $p = .016$, $\eta^2 = .06$, and for recipient B, $F(1, 95) = 11.059$, $p = .001$, $\eta^2 = .10$. Either way, treatment of neither recipient was privileged in determining punishment.

As shown in Figures 4a and 4c (see also Table S9 in the Supplemental Material), a similar (though mirror-reversed)

Fig. 3. Punisher spending in Study 2 in (a) the two-recipients condition and (b) the self-and-other condition. In (a), the mean amount punishers spent is shown as a function of how dictators treated recipients B and A, respectively. In (b), the mean amount punishers spent is shown as a function of how dictators treated the recipient and the punisher, respectively. Treatment was indexed by how much of their total stake dictators gave to each subject. Error bars indicate ±1 *SEM*.

pattern obtained for inferences: Punishers' inferences about how much dictators valued other subjects decreased with each act of poor treatment, whether it was directed at recipient A or B. This was true whether punishers were inferring valuation of themselves or recipient A.

Personal treatment trumps treatment of third parties. When punishers have access to information about personal treatment, do they stop using information about poor treatment of third parties? We tested this question using the self-and-other condition. A repeated measures ANOVA showed that the pattern of punishment was driven by whether punishers were themselves treated poorly, $F(1, 97) = 23.306$, $p < .001$, $\eta^2 = .194$ (Fig. 3b; see also Table S10 in the Supplemental Material). This effect was seven times larger than the comparable effect of poor treatment of the recipient, which did not reach significance, $F(1, 97) = 2.664$, $p = .106$, $\eta^2 = .027$. Moreover, there was no interaction between treatment of the punisher and treatment of the recipient in predicting punishment, $F(1, 97) = 0.005$, $p = .942$, $\eta^2 = .000$. The difference between the two patterns shown in Figure 3 reveals that once punishers know how dictators would actually treat them, they ignore how dictators treat other people. The deterrence hypothesis was supported; the social-benefits hypothesis was not.

How do inferences about valuation correlate with punishment? When the only information punishers had was how the dictator treated other people, they used this third-party cue to judge how highly the dictator valued them personally. As a result, punishers' inferences about dictators' valuations of themselves and of other people never diverged in these conditions. This was true in Study 1 and in the one-recipient and two-recipients conditions of Study 2.

In contrast, in the self-and-other condition of Study 2, when punishers had information about how dictators treated them personally, their inferences about how they themselves and the recipient were valued diverged: Punishers' inferences about how dictators valued them depended only on how they were personally treated; their inferences about how dictators valued recipients depended only on how recipients were treated (see Figs. 4b and 4d; see also Table S11 in the Supplemental Material).

The fact that punishers' inferences about how much dictators valued punishers and recipients sometimes diverged in Study 2 allowed us to test correlationally what we tested experimentally with the self-and-other condition: Are inferences about personal valuation a stronger predictor of punishment than inferences about third-party valuation? To test this, we used all three conditions from Study 2 and entered both variables as predictors into a hierarchical linear model (Table S12 in the Supplemental Material). While both variables could, in principle, predict punishment, only the inferences about personal treatment

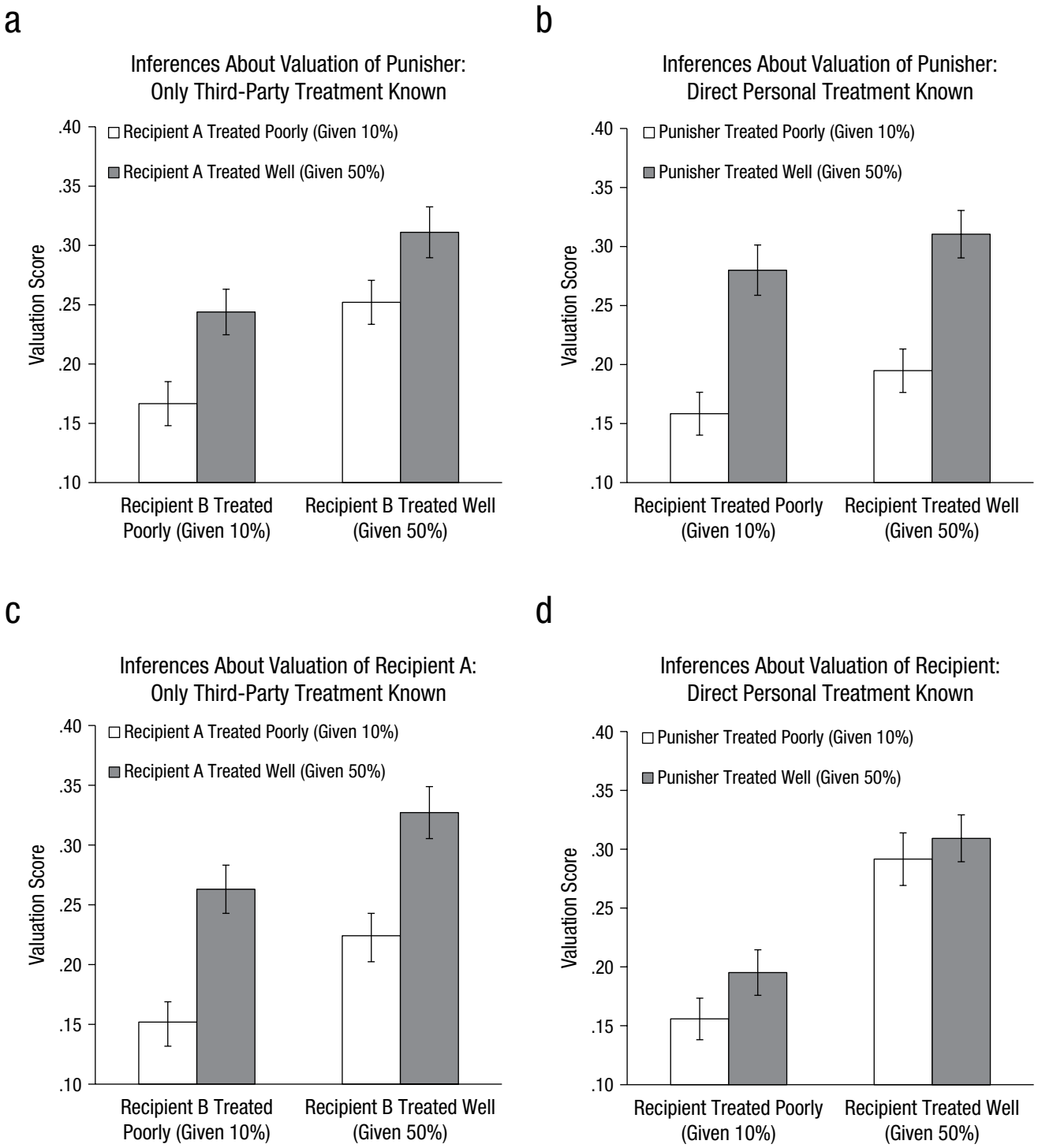

Fig. 4. Punishers' inferences in Study 2 about dictators' mean valuations in (a, c) the two-recipients condition and (b, d) the self-and-other condition. For the two-recipients condition, punishers' inferences about dictators' valuations of (a) the punisher and (c) recipient A are shown as a function of how dictators treated recipients B and A, respectively. For the self-and-other condition, punishers' inferences about dictators' valuations of (b) the punisher and (d) the recipient are shown as a function of how dictators treated the recipient and the punisher, respectively. Treatment was indexed by how much of their total stake dictators gave to each subject. Error bars indicate ±1 *SEM*.

did so—punisher: $\gamma = -0.12$, $t(293) = -3.01$, $p = .003$; recipient: $\gamma = 0.002$, $t(293) = 0.07$, $p > .250$. As the deterrence hypothesis predicts, punishers privileged information about personal treatment over third-party treatment.

Discussion

Humans and other animals regularly punish each other (Clutton-Brock & Parker, 1995). Such actions can recalibrate an offender's subsequent behavior and even lead the offender to become a valuable future partner (Krasnow et al., 2012). Thus, it is easy to see why natural selection would favor punishment mechanisms that respond to personal mistreatment. Less obvious is why natural selection would favor third-party interventions, including third-party punishment. Here, we reported tests of a novel account of human third-party punishment, the deterrence hypothesis.

Our approach assumes that the human decision-making architecture has been shaped by natural selection to solve evolutionarily recurrent, statistically common adaptive problems. One such adaptive problem is enforcing good treatment from other people who might otherwise be inclined to exploit you, your kin, or your allies. In real-world past environments, seeing mistreatment of someone else served as a predictive cue that you could also be mistreated by the offender. By punishing the offender when you witness someone else being mistreated, you can proactively deter the possibility that you or valued others will be mistreated in the future. When the only options available to laboratory subjects are to punish the offender or do nothing, the motivational pull of the inference that someone harming other people may harm you could be especially strong (Burton-Chellew & West, 2013; Pedersen, Kurzban, & McCullough, 2013).

We found that even anonymous third parties, alleged to be rationally disinterested, assume that dictators' treatment of a recipient predicts dictators' valuation of the third-party punisher. This inference, in turn, correlates with how much the punisher punishes the dictator. Moreover, in the absence of direct treatment of the punisher, punishers do not distinguish between how much dictators value recipients and how much dictators value punishers: Unless you know otherwise, assume that dictators will treat you the same way they will treat someone else. When information about personal treatment is available, however, this overrides information about treatment of anonymous other people. These findings are predicted and parsimoniously explained by the deterrence account. They are not consistent with a social-benefits account.

But could our results be due to another mechanism, signaling something besides one's potential response to personal mistreatment? By punishing third parties, people could signal their commitment to fairness, their cooperative disposition, or other relevant traits (Barclay, 2006; Baumard, André, & Sperber, 2013; Kurzban, DeScioli, & O'Brien, 2007). Punishment likely does sometimes function as such a signal, but this cannot explain the breadth of our results. Most critically, a cooperative-signaling account cannot explain why people punish less when they are personally treated well but a third party is treated poorly (Study 2). Any instance of poor treatment can be used to signal the punisher's commitment to fairness (holding constant the audience, direct costs of signaling, and so on). Indeed, punishing strongly when one is treated well but a third party is treated poorly would be a very good signal that the punisher has a disinterested commitment to fairness.

More generally, our results speak to the benefits of ecological-rationality approaches, with their focus on the cues that would have been available in past environments. For instance, other studies have shown that supposedly one-shot anonymous cooperation "irrationally" increases when subjects are exposed to stylized eyes or faces they know are not observers—more evidence that behavior is cue regulated (Haley & Fessler, 2005; Sparks & Barclay, 2013). The human mind embodies adaptive knowledge: knowledge that makes sense in the context of past environments, even if it appears irrational in ancestrally unrepresentative laboratory experiments.

Author Contributions

All authors contributed to the study design. Data were collected and analyzed by M. M. Krasnow and A. W. Delton, both of whom also drafted the manuscript. All authors made revisions and approved the final version of the manuscript for submission.

Acknowledgments

M. M. Krasnow and A. W. Delton contributed equally to this work.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by John Templeton Foundation Grant No. 29468 (http://www.templeton.org/). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplemental Material

Additional supporting information can be found at http://pss.sagepub.com/content/by/supplemental-data

Open Practices

All data have been made publicly available via Open Science Framework and can be accessed at https://osf.io/fhjyb. The complete Open Practices Disclosure for this article can be found

at http://pss.sagepub.com/content/by/supplemental-data. This article has received the badge for Open Data. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ and http://pss.sagepub.com/content/25/1/3.full.

Note

1. All *p* values reported are two-tailed.

References

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*, 325–344. doi:10.1016/j.evolhumbehav.2006.01.003

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral & Brain Sciences*, *36*, 59–78. doi:10.1017/S0140525X11002202

Brandts, J., & Charness, G. (2000). Hot vs. cold: Sequential responses and preference stability in experimental games. *Experimental Economics*, *2*, 227–238. doi:10.1023/A:1009962612354

Brosig, J., Weimann, J., & Yang, C.-L. (2003). The hot versus cold effect in a simple bargaining experiment. *Experimental Economics*, *6*, 75–90. doi:10.1023/A:1024204826499

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, *60*, 930–940. doi:10.1016/j.neuron.2008.10.016

Burton-Chellew, M. N., & West, S. A. (2013). Prosocial preferences do not explain human cooperation in public-goods games. *Proceedings of the National Academy of Sciences, USA*, *110*, 216–221. doi:10.1073/pnas.1210960110

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*, 209–216.

Delton, A. W. (2010). *A psychological calculus for welfare tradeoffs* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3427833)

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences, USA*, *108*, 13335–13340.

Delton, A. W., & Robertson, T. E. (2016). How the mind makes welfare tradeoffs: Evolution, computation, and emotion. *Current Opinion in Psychology*, *7*, 12–16.

Dunbar, R. I. M. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral & Brain Sciences*, *16*, 681–694. doi:10.1017/S0140525X00032325

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. doi:10.3758/BRM.41.4.1149

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87. doi:10.1016/S1090-5138(04)00005-4

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*, 1–25.

Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*, 245–256.

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition*, *27*, 733–763. doi:10.1521/soco.2009.27.5.733

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, *327*, 1480–1484. doi:10.1126/science.1182238

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*, 1767–1770. doi:10.1126/science.1127333

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*, 1362–1367. doi:10.1126/science.1153808

Jordan, J., McAuliffe, K., & Rand, D. (2015). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*. Advance online publication. doi:10.1007/s10683-015-9466-8

Kenrick, D. T., Griskevicius, V., Sundie, J. M., Li, N. P., Li, Y. J., & Neuberg, S. L. (2009). Deep rationality: The evolutionary economics of decision making. *Social Cognition*, *27*, 764–785. doi:10.1521/soco.2009.27.5.764

Kirkpatrick, M., Delton, A. W., Robertson, T. E., & de Wit, H. (2015). Prosocial effects of MDMA: A measure of generosity. *Journal of Psychopharmacology*, *29*, 661–668. doi:10.1177/0269881115573806

Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What are punishment and reputation for? *PLoS ONE*, *7*(9), Article e45662. doi:10.1371/journal.pone.0045662

Krasnow, M. M., & Delton, A. W. (in press). The sketch is blank: No evidence for an explanatory role for cultural group selection. *Behavioral & Brain Sciences*.

Krasnow, M. M., Delton, A. W., Tooby, J., & Cosmides, L. (2013). Meeting now suggests we will meet again: Implications for debates on the evolution of cooperation. *Scientific Reports*, *3*, Article 1747. doi:10.1038/srep01747

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*, 75–84. doi:10.1016/j.evolhumbehav.2006.06.001

Lieberman, D., Tooby, J., & Cosmides, L. (2007). The architecture of human kin detection. *Nature*, *44*, 727–731.

Locey, M. L., Jones, B. A., & Rachlin, H. (2011). Real and hypothetical rewards. *Judgment and Decision Making*, *6*, 552–564.

Marlowe, F. W., Berbesque, J. C., Barrett, C., Bolyanatz, A., Gurven, M., & Tracer, D. (2011). The 'spiteful' origins of human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *278*, 2159–2164. doi:10.1098/rspb.2010.2342

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *The American Economic Review*, *93*, 366–380.

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

Mishra, S. (2014). Decision-making under risk: Integrating perspectives from biology, economics, and psychology. *Personality and Social Psychology Review*, *18*, 280–307. doi:10.1177/1088868314530517

Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans *really* punish altruistically? A closer look. *Proceedings of the Royal Society B: Biological Sciences*, *280*, 20122723. doi:10.1098/rspb.2012.2723

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2012). To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior*, *33*, 682–695. doi:10.1016/j.evolhumbehav.2012.05.003

Raihani, N. J., Grutter, A. S., & Bshary, R. (2010). Punishers benefit from third-party punishment in fish. *Science*, *327*, 171. doi:10.1126/science.1183068

Roos, P., Gelfand, M., Nau, D., & Carr, R. (2014). High strength-of-ties and low mobility enable the evolution of third-party punishment. *Proceedings of the Royal Society B: Biological Sciences*, *281*, 20132661. doi:10.1098/rspb.2013.2661

Sparks, A., & Barclay, P. (2013). Eye images increase generosity, but not for long: The limited effect of a false cue. *Evolution and Human Behavior*, *34*, 317–322. doi:10.1016/j.evolhumbehav.2013.05.001

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, *16*, 167–171.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York, NY: Oxford University Press.

Vigilant bystanders deliver mugger to hands of justice. (2014, November 13). *Crown Heights News*. Retrieved from http://crownheights.info/crime/458745/vigilant-bystanders-deliver-mugger-to-hands-of-justice/

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, *20*, 415–432.

Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., & Cook, K. S. (2009). The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Sciences, USA*, *106*, 11520–11523. doi:10.1073/pnas.0900636106