Research Article

Psychological Science
2022, Vol. 33(5) 789–797

DOI: 10.1177/09567976211054786
www.psychologicalscience.org/PS

Material Benefits Crowd Out Moralistic Punishment

Tage S. Rai
Rady School of Management, University of California, San Diego

Abstract

Across four experiments with U.S.-based online participants ($N = 1{,}495$ adults), I found that paying people to engage in moralistic punishment reduces their willingness to do so. In an economic game with real stakes, providing a monetary bonus for engaging in third-party punishment of unfair offers nearly cut participants' willingness to do so in half. In judgments of hypothetical transgressions, participants viewed punishers who accepted payment as having worse character and rated the punishers' punitive actions as less morally acceptable. Willingness to engage in punishment was restored if participants were offered large enough payments or were told that punishment accompanied by payment still signals moral virtue. Data were consistent with a signal-corruption mechanism whereby payment interferes with the prosocial signal that moralistic punishment provides about a punisher's motives. These findings have implications for the cultural evolution of punishment and suggest that understanding perpetrators' sociomoral incentives is essential to implementing conflict-reduction policies.

Keywords

aggressive behavior, cooperation, decision making, evolutionary psychology, judgment, morality, motivation, punishment, rewards, violence

Received 1/2/21; Revision accepted 9/21/21

Psychological theories have often assumed that prohibitions against harming other people form the core of our moral sense (Schein & Gray, 2015; Turiel, 1983), and rational-choice models have often assumed that aggression occurs when its material benefits outweigh its costs (Cornish & Clarke, 2014; Freeman, 1999). However, there is growing recognition that much aggression is proximately motivated by moral and reputational concerns rather than *narrow material interests*, defined here as the first-order material costs and benefits directly tied to an exchange under standard economic assumptions (Gneezy & Rustichini, 2000). Ethnographic and historical analyses have found that perpetrators feel morally obligated to harm their victims (Fiske & Rai, 2014; Rai & Fiske, 2011), terrorists sacrifice their lives because they believe their cause is righteous (Atran, 2010), and nations go to war to defend their honor and standing (Lebow, 2010). In experiments, participants have experienced pleasure at their enemies' suffering (Cikara et al., 2014), and their willingness to harm has been better predicted by retributive desires than by deterrence (Carlsmith et al., 2002). These findings suggest that *moralistic aggression*, whereby perpetrators are motivated by broader moral and reputational concerns, may not be activated in response to direct material incentives in ways that are consistent with narrow material interests (Ginges, 2019).

Such dynamics have been explored most closely in studies of moralistic punishment. In economic games, third parties are willing to incur sizable monetary costs in order to remove resources from anonymous strangers who have morally transgressed in exchanges that the punishers have no direct material stake in (Fehr & Fischbacher, 2004). Game-theoretic models and empirical findings suggest that this behavior is supported by altruistic desires and reputational concerns on the part of punishers (Fehr & Gächter, 2002; Raihani & Bshary, 2015a). According to the reputational account, a punisher's willingness to incur monetary costs may signal

Corresponding Author:
Tage S. Rai, University of California, San Diego, Rady School of Management
Email: trai@ucsd.edu

trustworthiness and predispose observers to cooperate with them, stabilizing punishment (Barclay, 2006; Jordan et al., 2016; Nelissen, 2008; Raihani & Bshary, 2015b). In these studies, punishers were not required to incur direct monetary benefits in order to engage in punishment because doing so might introduce second-order cooperation challenges and incentivize short-term, self-serving punishment. Implicit in these design choices, however, is the assumption that providing material compensation as payment for punishment should simply increase willingness to engage in punishment compared with receiving no payment at all because payment serves a punisher's immediate material self-interest.

In this research, I experimentally investigated the impact of introducing monetary compensation for third-party punishment. Contrary to standard economic assumptions, my hypothesis was that under certain conditions, the introduction of material compensation as payment for punishment may actually decrease willingness to engage in third-party punishment of moral transgressions. This is because in the context of moralistic punishment, the absence of payment or other material benefits sends a clear signal to the punisher, the transgressor, and observers that the punisher is motivated by moral sentiments, which enhances the punisher's self-image and reputation (Raihani & Bshary, 2015a; Sarin et al., 2021). Accepting payment may interfere with this moral signal and send a negative signal (to other individuals as well as to oneself) about the punisher's intentions and character. According to this *moral-signal-corruption* hypothesis, when nonmaterial moral-signaling costs outweigh the narrow material benefits of accepting payment, then willingness to engage in punishment will decrease following payment.

Offers of payment have previously been found to reduce willingness to engage in helping behaviors such as aiding a fellow student (Deci, 1971), donating blood (Mellström & Johannesson, 2008), giving to charity (Eckel et al., 2005), and persuading other people to make donations (Barasch et al., 2016). Payment also appears to reduce support for peace (Ginges et al., 2007) and to offend recipients of "condolence payments" following collateral damage in war (Thomas, 2020). The present experiments extended these "crowding-out" and sociomoral-signaling effects (Bénabou & Tirole, 2006; Frey & Jegen, 2001) into the domain of aggression. Aggression is attractive from a theoretical standpoint because unlike helping behavior and conflict resolution, as described above, harming other people is often characterized as antisocial behavior that is extrinsically motivated and emits negative social signals (Coie & Dodge, 1998; Tedeschi & Felson, 1994), blunting traditional crowding-out mechanisms. More broadly, whereas recent literature has identified the moral motives underlying some forms of aggression, research exploring the links between moralistic aggression and broader psychological dynamics is still in its early stages and may challenge hypotheses that are based on models that assume that aggression is always antisocial in nature (Rai, 2019; Rai et al., 2017).

Statement of Relevance

Incentivizing a behavior should make it more likely to occur, but sometimes incentives backfire. One situation in which this may occur is the context of moral transgressions. To understand why, in the present research, I gave participants monetary rewards for punishing moral transgressions by other people. I found that paying people to engage in punishment actually reduced their willingness to do so. One reason for this may be that monetary rewards are incompatible with punishers' motives. By this reasoning, people engage in punishment in order to signal that they are good people themselves, but receiving compensation makes them seem like they are driven by money rather than justice. We may avoid this by paying punishers more money to overcome their moral qualms or by reframing payment as morally affirming. These findings highlight the importance of moral narratives surrounding punishment. They also suggest that conflict-reduction policies that emphasize material incentives may backfire if they fail to account for the moral motives that drive aggression.

General Method

To achieve requisite sample sizes, I recruited participants for all experiments via the Internet and compensated them with $0.20 after they completed a questionnaire administered through Amazon's Mechanical Turk (MTurk). It has been found that data collected through MTurk are as reliable as data gathered through traditional methods (Buhrmester et al., 2011). Duplicate responses that came from the same Internet protocol (IP) address were automatically eliminated to reduce the likelihood of participants completing the experiment multiple times. All research was approved by the university's institutional review board. Informed consent was obtained from all participants prior to participation. All materials and exact wordings of items can be found in the Supplemental Material available online.

Participants in all experiments reported demographic information, including their gender and ethnicity. All participants were from the United States. Of the 1,895 participants, 976 were men and 1,323 were White. No consistent meaningful interactions were predicted or found between the demographic variables and analyses of interest, so those analyses are not reported. Participants in all experiments were asked to guess the hypotheses of the study, but none succeeded. Each participant was assigned to condition randomly, and all experiments employed between-subjects designs. I sought 100 participants per cell in Experiment 1 because the design had two conditions. Experiments 2, 3a, and 3b each had three conditions, and therefore, 200 participants were sought per condition given the greater sample sizes required to assess interaction effects.

Experiment 1

Method

In Experiment 1, online participants were each presented with one of two versions of an incentive-compatible, binary-choice third-party punishment game. In this game, Player 1 was provided with an initial endowment of $20 from which they could allocate any portion to Player 2. Participants who completed the game as the third-party punisher (Player 3, $N = 196$) were the focus of my analysis. They were given the opportunity to incur a cost, which consisted of completing a boring arithmetic task, in order to remove $5 from Player 1 after learning how much money Player 1 allocated to Player 2. Any money removed from Player 1 was not reallocated to the participant or to Player 2; it was simply eliminated.

In the no-payment condition, participants had to make a single, binding yes/no choice about whether to complete the boring task in order to remove $5 from Player 1 if that player kept the entire endowment. This design was chosen to simplify the task for the online environment. In the payment condition, participants were presented with the same choice, but they were told that in addition to completing the boring task, they would also receive a small bonus payment of 1¢ if they chose to remove money from Player 1. Participants were informed that the experiment used a lottery system, whereby all choices were binding and approximately 1 in 20 participants would be randomly selected to have their choices played out for real stakes, a method that has been validated in prior research (Charness et al., 2016).

Note that the 1¢ bonus payment was low in absolute terms but not relative to the 20¢ participant fee. Thus, if participants in the payment condition were motivated by narrow material interests only, then the addition of any monetary bonus should not reduce their willingness to engage in punishment, compared with participants in the no-payment condition. Thus, the 1¢ bonus payment provided a strong proof of concept for the moral-signal-corruption hypothesis. Nevertheless, I explored the effects of increasing payment schemes in Experiment 3a.

Also, in a standard third-party economic game, participants playing as Player 3 pay monetarily to remove some portion of Player 1's earnings. However, because this experiment examined the effects of providing participants with a bonus monetary payment, the cost that participants paid to engage in punishment needed to be constituted in a different medium (i.e., the boring arithmetic task). Otherwise, the experimental manipulation of providing a monetary bonus would have simply equated to asking participants to pay a reduced cost to engage in punishment, rather than paying a cost and receiving a benefit. Participants were told that the arithmetic task would take approximately 10 min and that it was not fun or challenging but was meant to simulate "monotonous busy work." Pilot testing confirmed that participants found the prospect of the arithmetic task aversive.

Results

A χ^2 test revealed that participants were significantly more likely to engage in punishment when no payment was offered, compared with when they were offered a small payment, $\chi^2(1, N = 196) = 22.17$, $p < .001$, $\varphi = .34$. Overall, 71% (71/100) of participants in the no-payment condition decided to engage in punishment, a finding that is comparable with results of prior work using a standard third-party punishment game design. In contrast, only 38% (36/96) of participants engaged in punishment in the payment condition (see Fig. 1). Thus, paying participants to engage in punishment nearly cut their willingness to do so in half. The magnitude of this reduction is sizable and roughly equivalent to that achieved by tripling the cost of engaging in punishment in the standard version of the game (Egas & Riedl, 2008).

Experiment 2

Method

Experiment 1 demonstrated behaviorally that paying participants to engage in punishment reduces their willingness to do so. To explore and isolate potential causal mechanisms, I asked online participants in Experiment 2 ($N = 566$) to judge moral transgressions described in hypothetical vignettes. Participants adopted the perspective of an observer of moralistic punishment. Specifically,

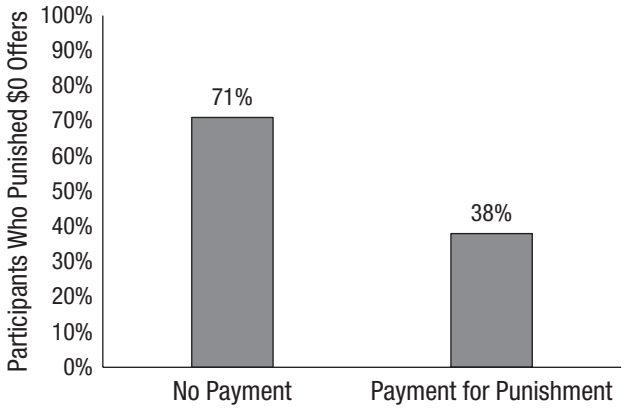

Fig. 1. Percentage of participants who committed to punishing Player 1 for keeping the entire endowment, separately for conditions in which participants received no payment and payment for engaging in punishment (Experiment 1).

they were asked to imagine that a person had decided to break a man's thumb as punishment for recruiting young women into prostitution. Depending on condition, participants were asked to imagine that the person had received no monetary payment, a small monetary payment (3¢), or a large monetary payment ($300) for physically harming the man. If there is a moral-signaling cost to accepting payment, then moral judgments should be more negative in the payment conditions. Varying the amount of payment also tested whether signaling costs are sensitive to levels of payment.

After reading the vignette, participants in each condition were asked to report (a) how morally acceptable breaking the man's thumb was (ranging from 1, *definitely morally wrong*, to 9, *definitely morally right*); (b) the extent to which the actor's primary motivation for doing it was clear (ranging from 1, *motivation is not clear*, to 9, *motivation is extremely clear*); (c) how pathetic it was (ranging from 1, *not pathetic*, to 9, *extremely pathetic*), which constituted my measure of moral character (reverse-scored so that higher values indicate superior moral character); and (d) how financially beneficial it was (ranging from 1, *not financially beneficial*, to 9, *extremely financially beneficial*). The last item acted as a manipulation check to ensure that participants were sensitive to the enhanced material benefits of the large payment. Not all participants answered all items, which resulted in some unequal cells.

Results

A one-way analysis of variance (ANOVA) found reliable differences across conditions in judgments of the moral acceptability of punishment, $F(2, 561) = 6.10$, $p = .002$, $\eta^2 = .021$; moral character, $F(2, 558) = 16.03$, $p < .001$, $\eta^2 = .054$; and the clarity of the punisher's motives, $F(2, 563) = 6.54$, $p = .002$, $\eta^2 = .023$ (see Fig. 2). Post hoc comparisons found that for moral acceptability of punishment, participants felt that engaging in punishment for no money ($n = 188$, $M = 4.60$, $SD = 2.83$) was more acceptable than doing so for a small payment ($n = 186$, $M = 3.69$, $SD = 2.91$), $t(372) = 3.04$, $p = .003$, or a large payment ($n = 190$, $M = 3.73$, $SD = 2.83$), $t(376) = 3.02$,

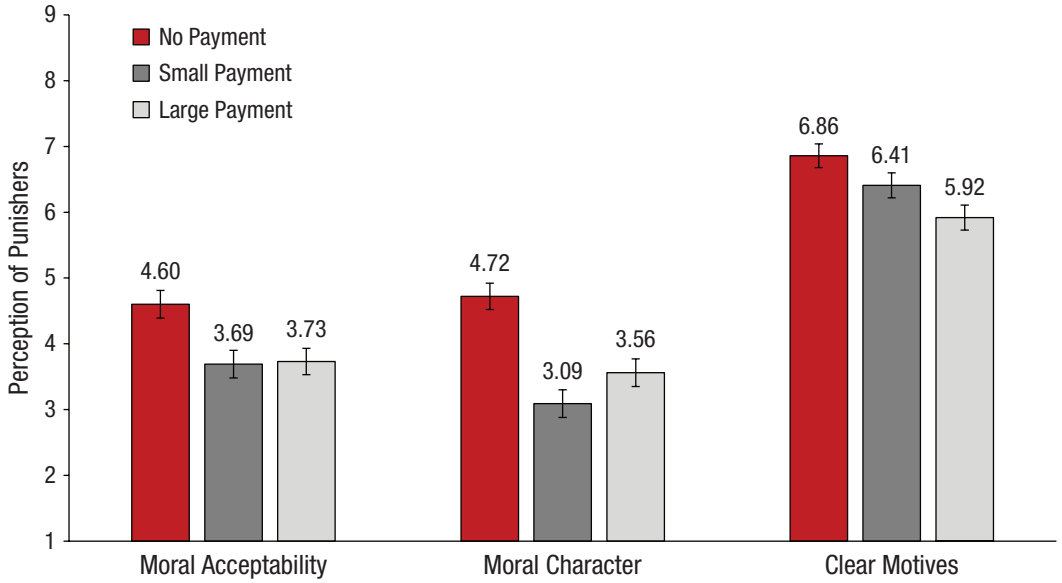

Fig. 2. Average rating of the moral acceptability of punishment and the punisher's moral character and motives in each payment condition (Experiment 2). Error bars represent standard errors.

$p = .003$. No significant difference was found in moral acceptability between the small-payment and large-payment conditions ($p = .911$). For moral character, participants felt that engaging in punishment for no money ($n = 186$, $M = 4.72$, $SD = 2.77$) was indicative of superior character, compared with doing so for a small payment ($n = 188$, $M = 3.09$, $SD = 2.91$), $t(372) = 5.55$, $p < .001$, or a large payment ($n = 187$, $M = 3.56$, $SD = 2.90$), $t(371) = 3.93$, $p < .001$. No significant difference was found between the small-payment and large-payment conditions ($p = .113$). For punisher's motives, participants felt that someone who engages in punishment for no payment ($n = 188$, $M = 6.86$, $SD = 2.43$) had clearer motives than someone who does so for a large payment ($n = 190$, $M = 5.92$, $SD = 2.59$), $t(376) = 3.66$, $p < .001$. No significant difference was found between the small-payment condition ($n = 188$, $M = 6.41$, $SD = 2.60$) and either the large-payment condition ($p = .065$) or the no-payment condition ($p = .083$). Combining the small- and large-payment conditions for punisher's motives ($M = 6.16$, $SD = 2.59$), I found that punishers who received no payment were seen as having clearer motives than punishers who received payment, $t(564) = 3.08$, $p = .002$.

Lastly, a manipulation check confirmed that participants perceived large payments as more incentivizing ($n = 190$, $M = 5.68$, $SD = 2.37$) than no payments ($n = 188$, $M = 2.44$, $SD = 2.19$) and small payments ($n = 188$, $M = 2.24$, $SD = 2.32$), $F(2, 563) = 134.01$, $p < .001$, $\eta^2 = .323$.

These findings suggest that payment corrupts the prosocial signal provided by moralistic punishment. Participants viewed punishment without payment as more acceptable than punishment with payment and rated the punisher as having superior character and acting with clearer motives when no payment was involved. The findings also suggest that moral-signaling costs are relatively fixed and that accepting any payment similarly contaminates the person who accepts it regardless of the size of payment, as there were no discernable differences in judgment across the small- and large-payment conditions, even though participants were sensitive to the material advantage of larger payments.

Because the benefits of payment increase with its quantity, it appears that individuals who engage in punishment for low compensation experience the worst of both worlds. Whereas people who engage in punishment for no compensation are seen as behaving more ethically, possessing superior moral character, and being driven by clearer motives than those who engage in punishment for high compensation, those who engage in punishment for low compensation do not receive any moral-signaling benefits over those who do so for high compensation, nor do they experience the material benefits of engaging in punishment for high compensation. This dynamic suggests two routes to restoring punishment in the presence of payment. One strategy is to compensate for moral-signaling costs by increasing material compensation. If moral-signaling costs do not increase with larger payment schemes, then eventually the material benefits of accepting payment should counterbalance the moral-signaling costs incurred by accepting payment. The other strategy is to repair the damage to the moral signal by influencing how the punisher morally construes their actions. If the punisher no longer believes that accepting payment corrupts the moral signal of punishment, then accepting payment should not reduce punishment. I explored these possibilities in Experiments 3a and 3b.

Experiment 3a

Method

If the moral-signaling cost of accepting payment is relatively fixed, as suggested by Experiment 2, then the relationship between payment and punishment should be nonmonotonic, whereby punishment drops at low levels of payment but recovers at high levels of payment. In Experiment 3a, participants ($N = 560$) were asked to imagine that they work at a firm and have evidence to prove that a coworker is engaging in insider trading that has earned him $100,000, and which if brought to light would result in the coworker's firing. Participants were asked to report how willing they would be to punish the coworker by reporting their actions (on a 7-point scale ranging from 1, *not willing at all*, to 7, *completely willing*). Depending on condition, willingness to punish was accompanied by no payment, a small payment, or a large payment. In the no-payment condition, participants were not given any monetary incentive to engage in punishment. In the low-payment condition, participants were told to imagine that they would receive $50 to carry out the punitive action. In the high-payment condition, participants were told to imagine that they would receive $50,000 to carry out the punitive action. For exploratory purposes, I also asked participants to rate how morally angry they felt toward the man (on a 9-point scale ranging from 1, *not angry*, to 9, *extremely angry*) and the extent to which they felt they would personally benefit from engaging in punishment (on a 9-point scale ranging from 1, *no personal benefit at all*, to 9, *large personal benefit*; see the Supplemental Material).

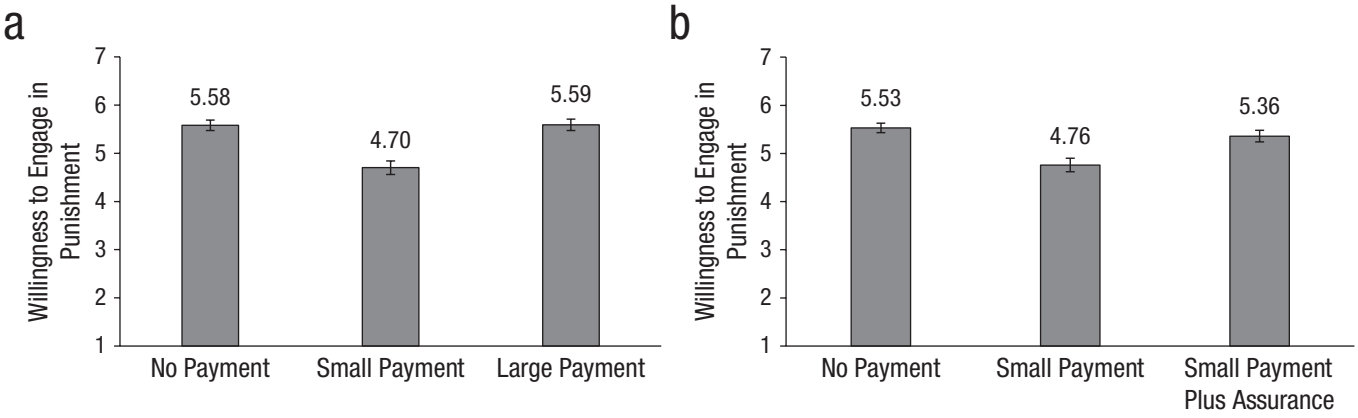

Fig. 3. Average rating of willingness to engage in punishment in each payment condition, separately for Experiments 3a (a) and 3b (b). Error bars represent standard errors.

Results

A one-way ANOVA revealed a significant effect of payment on willingness to engage in punishment, $F(2, 557) = 16.59$, $p < .001$, $\eta^2 = .056$. Post hoc comparisons revealed that participants expressed less willingness to engage in punishment when they received a small payment ($n = 193$, $M = 4.70$, $SD = 2.00$) than when they received no payment ($n = 187$, $M = 5.58$, $SD = 1.63$), $t(378) = 4.86$, $p < .001$, which conceptually replicated the findings from the economic game in Experiment 1. Crucially, participants also expressed less willingness to engage in punishment when they received a small payment than when they received a large payment ($n = 180$, $M = 5.59$, $SD = 1.77$), $t(371) = 4.66$, $p < .001$. No significant difference in willingness to engage in punishment was found between participants who received no payment and those who received a large payment ($p = .971$), which indicates that high levels of payment compensate for moral-signaling costs and restore punishment to baseline levels (see Fig. 3a).

Experiment 3b

Method

If reduced punishment following payment is driven by signal corruption, then payment should not reduce punishment if participants are made to believe that their actions still signal that they are engaging in punishment for moral reasons. In Experiment 3b, participants ($N = 573$) in the first two conditions were presented with either the no-payment or small-payment scenarios from Experiment 3a. A third group of participants was assigned to a small-payment-plus-assurance condition. The assurance condition was identical to the small-payment condition except that participants were also given the following information: "In our previous studies, most participants tell us that someone who turns in a partner for illegal insider trading in exchange for $50 is doing so because they feel the partner deserves it, not for the money." Participants then answered the same items as in Experiment 3a.

Results

A one-way ANOVA revealed a significant effect of condition on willingness to engage in punishment, $F(2, 570) = 10.89$, $p < .001$, $\eta^2 = .037$. Post hoc comparisons revealed that participants expressed reduced willingness to engage in punishment when they received a small payment ($n = 191$, $M = 4.76$, $SD = 1.96$) than when they received no payment ($n = 193$, $M = 5.53$, $SD = 1.45$), $t(382) = 4.38$, $p < .001$, which replicated the findings from Experiment 3a and conceptually replicated the findings from the economic game in Experiment 1. Crucially, participants also expressed reduced willingness to engage in punishment when they received a small payment than when they received a small payment with a positive moral assurance ($n = 189$, $M = 5.36$, $SD = 1.64$), $t(378) = 3.24$, $p = .001$. No significant difference in willingness to engage in punishment was found between participants who received no payment and those who received a small payment with a positive moral assurance ($p = .288$), which indicates that positive moral assurances repair punishers' beliefs about the moral signal of their actions and restore punishment to baseline levels (see Fig. 3b).

In the Supplemental Material, I report additional analyses that replicated the findings from Experiments 3a and 3b using a modified version of the physical-harm vignette from Experiment 2. These findings indicate that these effects extend across very different types of transgressions and punishments.

Discussion

Punishment norms often focus on destroying material value for recipients of harm more so than on providing any immediate material value to punishers. This is surprising because standard models based on narrow material interests would predict that providing direct material compensation to punishers as payment should only increase their willingness to engage in punishment. However, the current findings suggest that paying third parties to engage in moralistic punishment can actually reduce their willingness to do so under certain circumstances. This is because payment for punishment appears to introduce countervailing signaling costs by making punishers and their actions seem less ethical, which must be combated through larger payments or moral assurances. Future research should investigate whether providing payment for punishment also breaks down cooperation in repeated interaction settings even if antisocial punishment is prohibited, either by reducing a third party's willingness to engage in punishment or by interfering with recipients' and observers' ability to infer the moralistic intent of punishment. Exploring these dynamics would provide a fuller understanding of how moral signaling may constrain the cultural evolution of directly incentivized third-party punishment.

To the extent that perpetrators are often motivated by moral sentiments (Fiske & Rai, 2014; Rai & Fiske, 2011), these findings suggest that strategies for corrupting the moral-signaling value of aggression may offer an alternative approach to reducing harm compared with policy interventions that rely solely on leveraging material costs and benefits, which may be inefficient or even counterproductive. However, the findings do not suggest that to reduce aggressive behavior and conflict in real-world settings, we should consider paying violent actors to harm other people. Violent actors and organizations already provide material benefits as compensation for moralistic punishment and aggression in some cases (e.g., salaries to police and military officers). Yet links between material compensation and harm are often minimized to observers, and violent actors may even be penalized when such links are made explicit, such as when the public condemns private security forces and mercenaries (Ramirez & Wood, 2019). Future research should investigate whether cultural norms operate to restrict or obfuscate material incentives to violent actors and how these dynamics affect the compensation that they demand. Further, we should more closely examine the narratives that violent actors use to reinforce the prosocial signals that they communicate through moralistic aggression that is interwoven with material incentives, particularly in field settings. Such an understanding may provide insight into when and how to strategically call attention to material motives that accompany moralistic aggression as a means to corrupt its intended prosocial signal.

Caution is warranted when generalizing from lab experiments to the real world. The present experiments examined anonymous third-party punishment of strangers in response to moral transgressions that participants were not directly involved in and for which there was no prior history. In contrast, much moralistic punishment, and moralistic aggression more broadly, is committed by second parties acting against known relations in rich historical contexts that directly affect them (Pedersen et al., 2018). In addition, samples were composed of U.S.-based online participants who may place greater emphasis on inferring the mental states of punishers (McNamara et al., 2019) and who may have little experience with severe levels of violence. The participants did have experience with more mild forms of everyday moralistic aggression, which suggests that the moral-signaling dynamics identified in these low-stakes experiments may generalize to many real-world contexts that participants in a U.S.-based cultural context are familiar with. Finally, whereas the experiments focused on cases in which there was relative homogeneity in moral motivation among participants, there are multiple stakeholders with competing motives in real-world settings, and often there is no consensus about whether a behavior represents a moral transgression deserving of punishment or a virtuous act to be praised (Fiske & Rai, 2014; Rai & Fiske, 2011).

The present data do not allow social-signaling and self-signaling channels to be distinguished in order to explain how payment crowds out moralistic punishment. Whereas Experiment 3b suggests a direct route for social-signaling information to influence behavior, it is also possible that social signals affect the self-signaling that participants engage in to interpret the morality of their own behavior (Ariely et al., 2009; Bénabou & Tirole, 2011). Competing explanations for the results also include the possibility that the introduction of payment interferes with intrinsic motives to engage in punishment (Deci et al., 1999) as well as the possibility that the introduction of payment leads participants to infer greater negative utilities to punishment that must be compensated for (Bowles & Polania-Reyes, 2012; Gneezy et al., 2011). However, exploratory analyses from Experiments 3a and 3b (see the Supplemental Material) indicate that offers of payment do not reduce feelings of moral anger, which suggests that the results are not strictly due to a reduction in intrinsic motivation in the presence of payment. Meanwhile, participants also reported that punishment in exchange for a large

payment, which restores punishment to baseline levels, is most beneficial to the punisher, which suggests that participants were not using payment to infer additional negative utility associated with punishment. Still, further research is necessary to more fully test these and other competing hypotheses and to investigate whether crowding out of moralistic punishment and aggression relies on similar or different causal mechanisms than crowding out of helping behaviors.

Transparency

Action Editor: Eddie Harmon-Jones
Editor: Patricia J. Bauer
Author Contributions

T. S. Rai is the sole author of this article and is responsible for its content.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

Data for this study have not been made publicly available but can be requested from the author via email (trai@ucsd.edu). The design and analysis plans for the experiments were not preregistered.

Acknowledgments

I thank David Tannenbaum and Fiery Cushman for helpful discussions and reviews of earlier drafts of this article as well as my parents for watching the children while I revised the manuscript during the pandemic.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976211054786

References

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544–555.

Atran, S. (2010). *Talking to the enemy: Violent extremism, sacred values, and what it means to be human*. Penguin.

Barasch, A., Berman, J. Z., & Small, D. A. (2016). When payment undermines the pitch: On the persuasiveness of pure motives in fund-raising. *Psychological Science*, *27*(10), 1388–1397. https://doi.org/10.1177/0956797616638841

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652–1678.

Bénabou, R., & Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, *126*(2), 805–855.

Bowles, S., & Polania-Reyes, S. (2012). Economic incentives and social preferences: Substitutes or complements? *Journal of Economic Literature*, *50*(2), 368–425.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284–299.

Charness, G., Gneezy, U., & Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, *131*, 141–150.

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology*, *55*, 110–125.

Coie, J. D., & Dodge, K. A. (1998). Aggression and antisocial behavior. In W. Damon & N. Eisenberg (Eds.), *Handbook of child psychology: Social, emotional, and personality development* (pp. 779–862). John Wiley.

Cornish, D. B., & Clarke, R. V. (Eds.). (2014). *The reasoning criminal: Rational choice perspectives on offending*. Transaction Publishers.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, *18*(1), 105–115.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627–668.

Eckel, C. C., Grossman, P. J., & Johnston, R. M. (2005). An experimental test of the crowding out hypothesis. *Journal of Public Economics*, *89*(8), 1543–1560.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1637), 871–878.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Fiske, A. P., & Rai, T. S. (2014). *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press.

Freeman, R. B. (1999). The economics of crime. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3C, pp. 3529–3571). Elsevier.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*(5), 589–611.

Ginges, J. (2019). The moral logic of political violence. *Trends in Cognitive Sciences*, *23*(1), 1–3.

Ginges, J., Atran, S., Medin, D., & Shikaki, K. (2007). Sacred bounds on rational resolution of violent political conflict. *Proceedings of the National Academy of Sciences, USA*, *104*(18), 7357–7360.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 191–210.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, *115*(3), 791–810.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476.

Lebow, R. N. (2010). *Why nations fight: Past and future motives for war*. Cambridge University Press.

McNamara, R. A., Willard, A. K., Norenzayan, A., & Henrich, J. (2019). Weighing outcome vs. intent across societies: How cultural models of mind shape moral reasoning. *Cognition*, *182*, 95–108.

Mellström, C., & Johannesson, M. (2008). Crowding out in blood donation: Was Titmuss right? *Journal of the European Economic Association*, *6*(4), 845–863.

Nelissen, R. M. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, *29*(4), 242–248.

Pedersen, E. J., McAuliffe, W. H., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General*, *147*(4), 514–544.

Rai, T. S. (2019). Higher self-control predicts engagement in undesirable moralistic aggression. *Personality and Individual Differences*, *149*, 152–156.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57–75.

Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences, USA*, *114*(32), 8511–8516.

Raihani, N. J., & Bshary, R. (2015a). The reputation of punishers. *Trends in Ecology & Evolution*, *30*(2), 98–103.

Raihani, N. J., & Bshary, R. (2015b). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution*, *69*(4), 993–1003.

Ramirez, M. D., & Wood, R. M. (2019). Public attitudes toward private military companies: Insights from principal–agent theory. *Journal of Conflict Resolution*, *63*(6), 1433–1459.

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, Article 104544. https://doi.org/10.1016/j.cognition.2020.104544

Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*, *41*(8), 1147–1163.

Tedeschi, J., & Felson, R. (1994). *Violence, aggression, and coercive actions*. American Psychological Association.

Thomas, G. (2020). The costs of war: Condolence payments and the politics of killing civilians. *Review of International Studies*, *46*(1), 156–176.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge University Press.