MIT Open Access Articles

When do we punish people who don't?

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

As Published: 10.1016/J.COGNITION.2019.104040

Publisher: Elsevier BV

Persistent URL: https://hdl.handle.net/1721.1/136434

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-NonCommercial-NoDerivs License

Massachusetts Institute of Technology

When do we punish people who don't?

Justin W. Martin[1*&], Jillian J. Jordan[2*], David G. Rand[3,4], Fiery Cushman[5]

[1] Department of Psychology, Boston College, Chestnut Hill, MA 02467

[2] Department of Management and Organizations, Northwestern University, Chicago, IL 60208

[3] Department of Management Science, Massachusetts Institute of Technology, Cambridge, MA 02142

[4] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02142

[5] Department of Psychology, Harvard University, Cambridge, MA 02138

* These authors contributed equally to this work

& To whom correspondence should be addressed.

Boston College Department of Psychology,
140 Commonwealth Avenue
Chestnut Hill, MA 02467
E-mail: justinwmartin@gmail.com

Word count: 13,185

Keywords: Punishment, Cooperation, Norms, Adaptation

Electronic copy available at: https://ssrn.com/abstract=3080990

Abstract

People often punish norm violations. In what cases is such punishment viewed as normative—a behavior that we "should" or even "must" engage in? We approach this question by asking when people who fail to punish a norm violator are, themselves, punished. (For instance, a boss who fails to punish transgressive employees might, herself, be fired). We conducted experiments exploring the contexts in which higher-order punishment occurs, using both incentivized economic games and hypothetical vignettes describing everyday situations. We presented participants with cases in which an individual fails to punish a transgressor, either as a victim (second-party) or as an observer (third-party). Across studies, we consistently observed higher-order punishment of non-punishing *observers*. Higher-order punishment of non-punishing *victims*, however, was consistently weaker, and sometimes non-existent. These results demonstrate the selective application of higher-order punishment, provide a new perspective on the psychological mechanisms that support it, and provide some clues regarding its function.

Electronic copy available at: https://ssrn.com/abstract=3080990

1. **Introduction**

As humans, we often punish those who are antisocial: People who do not cooperate (Balliet & Van Lange, 2013; de Quervain et al., 2004; Fehr & Gächter, 2002; Gächter, Renner, & Sefton, 2008; Mathew & Boyd, 2011), who cause harm (Bone & Raihani, 2015; Buckholtz et al., 2008; Martin & Cushman, 2016b; McCullough, Kurzban, & Tabak, 2013; Morris, MacGlashan, Littman, & Cushman, 2017; Treadway et al., 2014) who behave unfairly (Cushman, Dreber, Wang, & Costa, 2009; FeldmanHall, Sokol-Hessner, Van Bavel, & Phelps, 2014; Henrich et al., 2010, 2006; Martin & Cushman, 2015; Mendoza, Lane, & Amodio, 2014) or who violate norms (Balafoutas & Nikiforakis, 2012; Carpenter & Matthews, 2009; Fehr & Fischbacher, 2004b). Such punishment can enforce prosociality and uphold norms, whether on behalf of ourselves or others (Balliet, Mulder, & Van Lange, 2011; Carpenter & Matthews, 2009; Fehr & Fischbacher, 2004b; Gaechter, 2014).

Furthermore, punishment is itself sometimes *normative*. In particular, in some contexts punishment is “injunctively” normative, such that people feel that others *should* punish, or feel socially pressured to punish themselves (Whitson, Wang, See, Baker, & Murnighan, 2015). For instance, a parent whose child hits another child on the playground might feel that they should punish their child, or are expected to; likewise with a boss whose employee engages in sexual harassment. Yet, we have relatively little understanding of the contexts in which punishment is considered injunctively normative.

In order to study this question, we take advantage of a hallmark property of injunctively normative behaviors: failure to perform them is sometimes punished. In

Electronic copy available at: https://ssrn.com/abstract=3080990

other words, one way that we know that punishment is sometimes seen as normative is that individuals who fail to punish are sometimes *themselves* punished. (For instance, a boss who does nothing about sexual harassment among her subordinates might herself be punished). Such “higher-order punishment” (HOP) has attracted substantial theoretical interest (e.g. Brandt, Hauert, & Sigmund, 2006; Fowler, 2005). However, relatively little empirical research has investigated the contexts in which people do—or do not—engage in HOP of non-punishers (but see Cinyabuguma, Page, & Putterman, 2006; Fu, Ji, Kamei, & Putterman, 2017; Kiyonari & Barclay, 2008). This is our focus.

Specifically, we draw on a key conceptual distinction between two kinds of first-order punishment (FOP; i.e., the punishment of norm violations other than the failure to punish, such as interpersonal harm, theft, non-cooperation, etc.). Specifically, in line with past work (Fehr & Fischbacher, 2004a; FeldmanHall et al., 2014; Gummerum & Chu, 2014; Leibbrandt & López-Pérez, 2012; Raihani & Bshary, 2015a; Zhou, Jiao, & Zhang, 2016), we distinguish between "second-party punishment" (2PP), in which the victim of a transgression personally punishes the perpetrator, and "third-party punishment" (3PP), in which an uninvolved individual punishes on behalf of the victim. We hypothesize that HOP may be more likely in contexts where a third party failed to punish, as compared to contexts where a victim failed to punish (i.e. as a second-party).

To bring out the intuition behind this hypothesis, consider a concrete example. Suppose that Janet participates in a local basketball league. She steals money from her teammate Tom, and another teammate, Lisa (a “third-party”), is the only witness. Despite what Lisa sees, she does nothing to retaliate against or even express her disapproval of Janet's behavior. Should Lisa be reprimanded for doing nothing to stick

Electronic copy available at: https://ssrn.com/abstract=3080990

up for Tom? If Lisa's teammates decide to punish her, this would constitute HOP in the context of third-party punishment. Our intuition is that Lisa’s teammates might consider her inaction counter-normative, and thus might respond with some (perhaps minimal) form of punishment (e.g., by verbally rebuking her).

Now consider a different scenario without Lisa. Again, Janet steals from Tom, but this time Tom himself is the witness. Although he is upset, Tom does nothing to punish her—no retaliation, not even a harsh word. Would he be reprimanded for not sticking up for *himself*? If Tom's teammates decide to punish him for failing to stand up for himself, this would constitute HOP in the context of second-party punishment. Here, our intuition is that Tom’s teammates may consider him a “pushover” or draw various negative character inferences about him. However, importantly, we expect that they would not see Tom’s inaction violating normative standards for behavior, and would be unlikely to rebuke him. To preview our results, we find substantial support for this basic intuition.

Relatively few past studies have experimentally investigated HOP. Most of these studies have focused on HOP in the context of the public goods game (PGG) (Cinyabuguma et al., 2006; Cinyabuguma, Page, & Putterman, 2004; Fu et al., 2017; Kiyonari & Barclay, 2008). These studies have tended to find little or no HOP in the PGG (Cinyabuguma et al., 2006; Fu et al., 2017; Kiyonari & Barclay, 2008). Importantly, however, they do not clearly indicate whether we should expect HOP in straightforward cases of second- versus third-party punishment. Punishment in the PGG can be thought of as a middle ground between second- and third-party punishment. When an individual fails to contribute to a public good, this choice hurts all members of the group. Thus, a group member who punishes is responding to an act that harmed them personally, but also one that harmed others. Therefore, punishment in the PGG is a hybrid between 2PP and 3PP, and investigating HOP in this context does not shed light on the extent to which HOP is applied to “pure” forms of 2PP and 3PP.

Here, we aim to provide a strong contrast between cases of second-party and third-party punishment, and investigate when punishment is seen as normative (indexed by HOP) in a more targeted manner. To this end, we seek convergent evidence from two experimental methods. First, we have people play structured economic games in which they have an opportunity to punish non-punishers at a cost to themselves. Second, we present other participants with hypothetical vignettes involving everyday acts of antisocial behavior and ask how much non-punishers should be punished. These methods have complementary strengths: The economic games investigate real behavior that is subject to monetary incentives, while the vignettes investigate HOP in concrete, ordinary situations.

2. Experiment 1

Experiment 1 tests for higher-order punishment (HOP) in the context of a multiplayer economic game (Fig. 1). In the first stage, a moral transgression can occur. Two participants each earn a small bonus for themselves. Then, one of them (the potential "perpetrator") is given the opportunity to add a small amount to their own bonus (increasing it by 1/3). However, to do so they must destroy the entire bonus of the second participant (the potential "victim"). Our analysis focuses selectively on

Electronic copy available at: https://ssrn.com/abstract=3080990

cases where the potential perpetrator *does* “steal” from the potential victim in this way, and a transgression does occur.

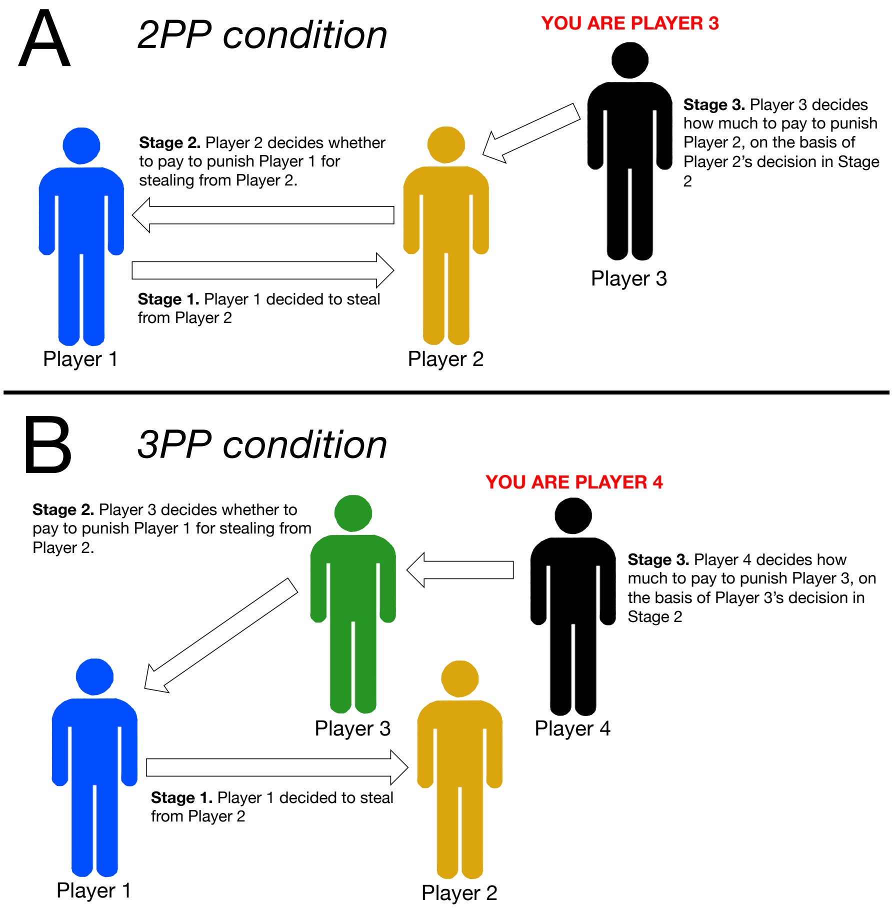
A
2PP condition
YOU ARE PLAYER 3
Stage 3. Player 3 decides how much to pay to punish Player 2, on the basis of Player 2’s decision in Stage 2
Stage 2. Player 2 decides whether to pay to punish Player 1 for stealing from Player 2.
Stage 1. Player 1 decided to steal from Player 2
Player 1
Player 2
Player 3
B
3PP condition
YOU ARE PLAYER 4
Stage 2. Player 3 decides whether to pay to punish Player 1 for stealing from Player 2.
Stage 3. Player 4 decides how much to pay to punish Player 3, on the basis of Player 3’s decision in Stage 2
Stage 1. Player 1 decided to steal from Player 2
Player 1
Player 2
Player 3
Player 4

Figure 1. The design of Experiment 1. Participants learned about a failure to punish a theft, either by the victim (2PP condition) or by an observer (3PP condition), and were then given the opportunity to engage in costly punishment of the non-punisher.

In the second stage, first-order punishment can occur. A “first-order punisher” participant is either the victim of the perpetrator’s transgression (2PP condition) or a third, impartial participant (3PP condition). In both conditions, the first-order punisher is endowed with some money and has the chance to sacrifice some of

Electronic copy available at: https://ssrn.com/abstract=3080990

this money to take money away from the perpetrator. Specifically, first-order punishers made a binary punishment decision. They decided whether to pay 1/5 of their endowment to remove the perpetrator's entire bonus (i.e., both the perpetrator's initial endowment and the amount that the perpetrator "stole").

Finally, in the third stage, higher-order punishment can occur. A "higher-order punisher" participant—whose behavior we focus on in this paper—learns about the first-order punisher's behavior, and can respond with higher-order punishment of the first-order punisher. Specifically, higher-order punishers were endowed with some money and made continuous punishment decisions. For every 1 unit they spent, 5 units were subtracted from the first-order punisher.

We predicted that in this game, participants would selectively punish non-punishers. In other words, we predicted more HOP when the first-order punisher failed to punish the perpetrator than when the first-order punisher chose to punish the perpetrator. Additionally, we predicted stronger punishment of observers who failed to punish perpetrators (in the 3PP condition) than of victims who fail to punish perpetrators (in the 2PP condition).

2.1 Exp. 1 Methods

Participants (*N* = 585) were recruited through Amazon Mechanical Turk to play an economic game. The design of our economic game is summarized in Figure 1. Participants interacted with other participants in a multi-stage game for real stakes. In stage 1, two participants (hereafter called the Perpetrator and the Victim) played a version of the Dictator Game (DG). Each earned $0.15 for performing a short task. Then, the Perpetrator was given the opportunity to steal the Victim's bonus. If the

Electronic copy available at: https://ssrn.com/abstract=3080990

Perpetrator did steal the bonus, they received an extra $0.05, while the Victim lost all $0.15. Only Perpetrators who did steal this bonus were included in Stage 2 of this study. In Stage 2, participants (hereafter "First-Order Punishers") were given the opportunity to engage in costly punishment of the Perpetrator: They were given $0.25 and told that they could pay $0.05 to reduce the Perpetrator’s bonus by $0.20 (leaving the Perpetrator with nothing). In Supplemental Experiment S1 punishment was instead costless for First-Order Punishers; overall, this design yielded the same pattern of results. For participants assigned to the second-party condition (*n* = 290), the First-Order Punisher was the initial Victim. For participants assigned to the third-party condition (*n* = 295), the First-Order Punisher was an impartial observer who was told all details from stage 1. In stage 3, the experimental participants of interest evaluated the decision of First-Order Punishers, either the Victim (second-party condition) or the Observer (third-party condition). Participants were given $0.25 and could engage in costly punishment of the First-Order Punisher, paying between 0 and 4 cents to reduce the First-Order Punisher’s payoff by $0.05 for each cent paid (up to a $0.20 reduction). This decision (how much participants punished First-Order Punishers in the second- and third-party conditions) was our variable of interest.

To maximize the amount of data collected per participant, the strategy method was used: Participants indicated how much they wished to sanction in the event that the First-Order Punisher decided to punish, and in the event that the First-Order Punisher decided to not punish. Participants were first given instructions for the economic game, then asked six comprehension questions about the payoff structure of the game, and then participated in the game. We excluded participants who did not answer all

Electronic copy available at: https://ssrn.com/abstract=3080990

comprehension questions correctly and participants who responded in less than 2.5 seconds for either question about imposing sanctions. Based on these exclusionary criteria, data from 179 participants (30.6%) were discarded, with 99 excluded from the second-party condition (34.1%) and 80 excluded from the third-party condition (27.1%). These proportions did not differ significantly (two-sample proportion test $X^2 = 3.07$, $n = 585$, $p = 0.08$). Including all subjects does not change the overall pattern of results. Participants then answered standard demographic questions as well as a question regarding their belief in whether their partner was real. Including this factor in analyses did not change the overall pattern of results. All procedures were approved by the Yale University Institutional Review Board.

For this experiment and all following experiments, data were principally analyzed using mixed-effects regression. Our primary analyses investigated the amount of punishment applied by higher-order punishers. In these analyses, fixed effects included overall condition (second- versus third-party, between-subjects) and whether or not punishment occurred (within-subjects), as well as their interaction. We included a random intercept for each subject. We used linear mixed effects regression implemented in R using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014). P-values were obtained for fixed effects using the Kenward-Roger approximation of degrees of freedom, implemented in lmerTest (Kuznetsova, Brockhoff, & Christensen, 2015) and pbkrtest (Halekoh & Højsgaard, 2014). We also did some analyses investigating the probability of enacting any higher-order punishment. In these analyses, we used logistic mixed effects regression, implemented in R using the lme4 package

Electronic copy available at: https://ssrn.com/abstract=3080990

(Bates et al., 2014), with the same model specification as above. Data and analysis scripts for all experiments can be found at https://osf.io/fhmnd/.

2.2 Exp 1. Results

2.2.1 Amount spent on HOP

We first analyze the amount of HOP assigned using linear mixed-effects regression. We take a model comparison approach, starting with a "full" model that includes all predictors and comparing to one without the interaction term, and then comparing this reduced model to models dropping each of the main effects. We find a significant interaction between condition (2PP vs. 3PP) and whether FOP did or did not occur (1000 sample bootstrap LRT $X^2(1) = 19.54$, $p < 0.001$, **β** = 0.13, SE = 0.03, 95% CI = 0.07 – 0.19). We also find a significant main effect of condition (1000 sample bootstrap LRT $X^2(1) = 15.23$, $p < 0.001$, **β** = 0.15, SE = 0.04, 95% CI = 0.08 – 0.23), and whether the FOP occurred or not (1000 sample bootstrap LRT $X^2(1) = 9.83$, $p < 0.005$, **β** = 0.09, SE = 0.03, 95% CI = 0.04 – 0.15). Given the presence of the interaction, we followed up the significant main effects with tests of the simple effects within each condition (2PP and 3PP).

First, looking at mean HOP in the 3PP condition, we observe greater HOP when FOP did not occur than when FOP did occur (Figure 2; Did not punish: $M = 0.61$, $SEM = 0.09$; Did punish: $M = 0.23$, $SEM = 0.05$; paired $t(214) = 4.24$, $p < 0.001$, Cohen's $d = 0.41$, 95% CI = 0.22 – 0.60). We find no significant effect, however, in the 2PP condition (Did not punish: $M = 0.10$, $SEM = 0.04$; Did punish: $M = 0.18$, $SEM = 0.05$; paired $t(190) = 1.90$, $p = 0.06$, Cohen's $d = 0.19$, 95% CI = -0.01 – 0.40). Further, greater HOP was levied when participants assessed third parties who failed to punish relative to participants who assessed second parties who failed to punish (Welch *t*(289.89) = 5.34, *p* < 0.001, Cohen's *d* = 0.53, 95% CI = 0.33 – 0.73) but no difference when they do punish (Welch *t*(403.97) = 0.61, *p* = 0.54, Cohen's *d* = 0.06, 95% CI = -0.13 – 0.26) Thus, we find that whether or not a first-order punisher chose to punish selfishness influences how much HOP they receive when they are a third-party, but not when they are a second-party; and that more HOP is assigned to third parties who fail to punish than second parties who fail to punish. And, we find the same pattern of results when FOP is costless (see Experiment S1).

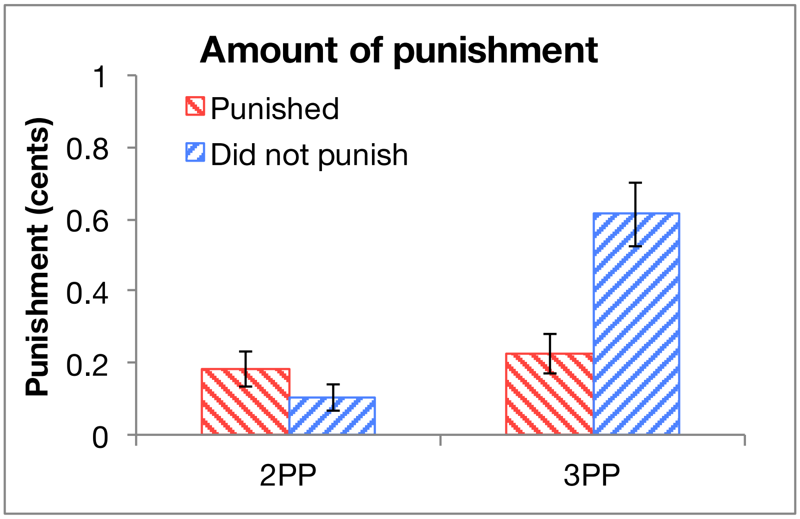

Figure 2. Results of Exp. 1. Participants were given the opportunity to punish a first-order punisher (second- vs. third-party) who either did or did not engage in costly punishment. Amount spent on punishment (out of a maximum of 4 cents) by participants in the second- or third-party condition, in cases where FOP either did or did not occur. Error bars are SEM.

2.2.2 Percentage of participants engaging in HOP

Our data allow us to look not just at the overall amount of HOP that occurs, but also the proportion of participants who engage in HOP. Because the mean amount of

Electronic copy available at: https://ssrn.com/abstract=3080990

HOP enacted could come from a minority of individuals, looking at the proportion of participants who engage in HOP in each condition provides a sense of how consistent or widespread preferences for HOP are. Here, we classified each participant as either not engaging in HOP (i.e. assigning \$0 in sanctions) or engaging in HOP (i.e. assigning any amount greater than \$0 in sanctions) and again analyzed results using mixed-effects regression.

Overall, we again find a significant interaction between whether or not punishment occurred and whether the participant was assessing a second- or third-party (1000 sample bootstrap LRT $X^2(1) = 36.9$, $p = 0.002$, Odds Ratio = 7.03, 95% CI = 3.27 – 18.04) as well as a main effect of condition (1000 sample bootstrap LRT $X^2(1) = 3.30$, $p = 0.001$, Odds Ratio = 1.60, 95% CI = 0.59 – 4.89) and whether or not FOP occurred (1000 sample bootstrap LRT $X^2(1) = 14.72$, $p = 0.01$, Odds Ratio = 1.31, 95% CI = 0.67 – 2.52). Looking at the proportion of participants engaging in HOP in the third-party condition, we find that more individuals punish when FOP did not occur than when FOP did occur (Did not punish: 22.3%, *SE* of proportion = 0.03; Did punish: 9.7%, *SEP* = 0.02; McNemar's $X^2(1, N = 215) = 9.80$, $p = 0.001$, Odds Ratio = 2.21, 95% CI of OR = 1.34 – 3.64). This is not the case for the second-party condition (Did not punish: 4.7%, *SEP* = 0.02; Did punish: 8.4%, *SEP* = 0.02; McNemar's X^2 $(1, N = 191) = 1.44$, $p = 0.23$, Odds Ratio = 1.37, 95% CI of OR = 0.81– 2.31). Finally, more participants engage in HOP of third parties who failed to punish than second parties who failed to punish (two-sample proportion test X^2 $(1, N = 406) = 24.6$, $p < 0.001$, Odds Ratio = 2.51, 95% CI of OR = 1.74 – 3.61). Thus, we find that whether or not FOP occurred influences how many people engage in HOP when they are a third-party, but not when

Electronic copy available at: https://ssrn.com/abstract=3080990

they are a second-party; and that more people engage in HOP of third parties who fail to punish than second parties who fail to punish.

2.2.3 Amount spent on HOP given engagement in HOP

Consistent with this difference in the proportion of individuals engaging in HOP between conditions, we find no difference between conditions in the amount spent on punishment given that one decided to punish, either when FOP did occur (2PP: *M* = 2.19, *SEM* = 0.31; 3PP: *M* = 2.33, *SEM* = 0.25; Welch two-sample *t*(31.4) = 0.37, *p* = 0.72, Cohen's *d* = 0.12, 95% CI = -0.55 – 0.8) or when FOP did not occur (2PP: *M* = 2.22, *SEM* = 0.36; 3PP: *M* = 2.75, *SEM* = 0.18; Welch two-sample *t*(12.1) = 1.3, *p* = 0.22, Cohen's *d* = 0.47, 95% CI = -0.26 – 1.21). We note, however, that the small number of participants involved in these comparisons necessitates caution in their interpretation.

2.3 Exp. 1 Discussion

In sum, we find that whether or not a first-order punisher chose to punish selfishness influences how much HOP they receive when they are a third-party, but not when they are a second-party; and that more HOP is assigned to third parties who fail to punish than second parties who fail to punish. Thus, Experiment 1 provides evidence for HOP, but specifically in the context of 3PP. Participants sacrificed their own money to punish third-party observers who failed to punish stealing (while punishment of third-parties who did punish stealing was very rare). We note that even in the context of 3PP, we observed a small absolute amount of HOP of non-punishers: 0.6 out of a possible 5 cents, or 12% of the maximum punishment that we could have observed. Moreover, only a minority of individuals (22%) punished third-party observers who failed to punish. Nonetheless, however, we observed significantly more HOP in the context of 3PP than 2PP, and thus our results reveal something meaningful about the psychology underlying HOP, and the contexts in which it occurs.

3. Experiment 2

The economic game employed in Experiment 1 has the virtue of measuring incentive-compatible behavior, but the drawback that it is highly abstract. Additionally, the incentives were quite low compared to many relevant real-world settings, and Experiment 1 employed the “strategy method”, which may have made HOP decisions less emotional (i.e., more “cold” than “hot”) than HOP decisions in everyday contexts. In Experiment 2 we aim to provide convergent evidence from participants' judgments of concrete vignettes. Although these vignettes are hypothetical, they contextualize phenomena of interest within familiar everyday settings. Specifically, the vignettes describe ordinary situations (e.g. reading a book at a coffee shop, having dinner) modeled on the structure of our economic game in Experiment 1. Thus, despite being hypothetical, the vignettes describe situations with much high stakes than then we implemented in our economic games. Additionally, by telling stories about concrete transgressions that have already occurred, the vignettes may evoke more emotional processing, and thus better approximate the psychology of HOP in daily life. In sum, our aim in Experiment 2 is to provide convergent evidence using a distinct methodology.

3.1 Exp 2. Methods

Participants (*N* = 899) read through six vignettes designed to closely match the structure of the economic game in Experiment 1. All vignettes involved a perpetrator who harmed a victim. As in the economic game, we varied two factors orthogonally:

Whether the potential punisher was a second- or third-party, and whether they did or did not engage in punishment, yielding a 2 x 2 design. In this and all subsequent studies, we varied both factors between-subjects, in an effort to avoid demand effects. The text of all vignettes was only minimally changed across the 4 conditions. In the second-party condition, the victim is present as or immediately after the harm takes place and has the opportunity to engage in punishment. In the third-party condition, a third-party (and not the victim) is present as or immediately after the harm takes place and has the opportunity to engage in punishment. In both of these conditions, the potential punisher either does (2PP *n* = 227; 3PP *n* = 223) or does not (2PP *n* = 228; 3PP *n* = 221) engage in punishment, and the perpetrator leaves the scene. The text of all cases can be found in the Supplemental Information.

Participants were asked how much the potential punisher should be punished, on a scale of 1 to 10 with anchors at 1 = “No punishment at all”, 2 = “Minimal punishment”, 6 = “Moderate punishment” and 10 = “Extreme punishment”. Because the response indicating zero punishment was actually coded as “1”, our results have been re-baselined, by subtracting 1 from all responses. Participants were randomly assigned to one of the four conditions and read all vignettes for the experiment in that condition. After the final trial, participants completed attention check questions, for use in assessing data quality. Participants were excluded based on responses to questions regarding their attentiveness to the study, if their average reaction time across the six vignettes was less than 8 seconds and if they reported not being a native speaker of English. Based on these criteria, data from 123 participants (out of 899, 13.7%; 2PP, did punish = 32; 2PP, did not punish = 34; 3PP, did punish = 26; 3PP, did not punish =

Electronic copy available at: https://ssrn.com/abstract=3080990

31) were discarded. Percentage excluded was nearly identical across all four conditions (2PP, did punish: 14.1%; 2PP, did not punish: 14.9%; 3PP, did punish: 11.7%; 3PP, did not punish: 14.0%). Including all subjects does not change the overall pattern of results. All procedures were approved by the Harvard University Committee on the Use of Human Subjects.

Data analysis was carried out as specified in Experiment 1, with two exceptions. First, whether or not first-order punishment (FOP) occurred is now a between-subjects variable. Second, because participants were exposed to multiple trials with different vignette contexts, we now include a random intercept for vignette.

3.2 Exp 2. Results

3.2.1 Endorsement of HOP

We first analyze the amount of higher-order punishment (HOP) endorsed using linear mixed-effects regression. We find a significant interaction between condition (2PP vs. 3PP) and whether or not the protagonist punished (1000 sample bootstrap LRT $X^2(1) = 34.39$, $p < 0.001$, $\boldsymbol{\beta} = 0.17$, SE = 0.03, 95% CI = 0.11 – 0.22). We also find a significant main effect of condition (1000 sample bootstrap LRT $X^2(1) = 60.65$, $p < 0.001$, $\boldsymbol{\beta} = 0.23$, SE = 0.03, 95% CI = 0.17 – 0.29), and whether FOP occurred or not (1000 sample bootstrap LRT $X^2(1) = 61.38$, $p < 0.001$, $\boldsymbol{\beta} = 0.23$, SE = 0.03, 95% CI = 0.18 – 0.29). Given the presence of the interaction, we followed up the significant main effects with tests of the simple effects within each condition (2PP and 3PP).

First, looking at HOP in the 3PP condition, we observe more endorsement of HOP of non-punishing protagonists than of punishing protagonists (Figure 3; Did not punish: $M = 1.41$, $SEM = 0.13$; Did punish: $M = 0.29$, $SEM = 0.07$; Welch $t(291.9) =$

7.83, *p* < 0.001, Cohen's *d* = 0.8, 95% CI = 0.59 – 1.0). We find the same pattern in the 2PP condition (Did not punish: *M* = 0.29, *SEM* = 0.06; Did punish: *M* = 0.12, *SEM* = 0.04; Welch *t*(337.14) = 2.45 *p* = 0.014, Cohen's *d* = 0.25, 95% CI = 0.05 – 0.45), though we observe a much smaller effect. Finally, more HOP was observed when participants assessed third parties relative to participants who assessed second parties, either when they failed to punish (Welch *t*(272.68) = 8.02, *p* < 0.001, Cohen's *d* = 0.82, 95% CI = 0.61 – 1.03) or when they did punish (Welch *t*(319.07) = 2.26, *p* = 0.024, Cohen's *d* = 0.23, 95% CI = 0.03 – 0.43).

3.2.2 Endorsement of HOP relative to baseline

In a separate study, we also examined levels of HOP endorsed in the context of 3PP and 2PP (focusing only on cases in which either 3PP or 2PP did *not* occur), relative to a neutral baseline condition (in which subjects evaluated an entirely uninvolved individual). This allowed us to better interpret absolute levels of HOP of non-punishers that subjects endorsed, both in the context of 3PP and 2PP. We replicated our prior results, and also found that participants endorsed more punishment of non-punishers (both in the 3PP and 2PP conditions) than of entirely uninvolved individuals (in the neutral baseline condition) (see Supplemental Experiment S2). Thus, we find that subjects endorse above-baseline HOP of non-punishers in both 3PP and 2PP conditions, though we observe an effect 4 times as large for third parties; and we find that subjects endorse more HOP of third parties who fail to punish than of second parties who fail to punish.

3.2.3 Proportion of trials on which HOP was endorsed

Electronic copy available at: https://ssrn.com/abstract=3080990

As in Experiment 1, we analyze not just at the amount of HOP that participants endorsed, but also the proportion of trials in which participants endorsed any HOP. We classified each trial as one on which the participant either did not endorse HOP (i.e. punishment = 0) or endorsed HOP (i.e. punishment > 0) and again analyzed results using mixed-effects regression.

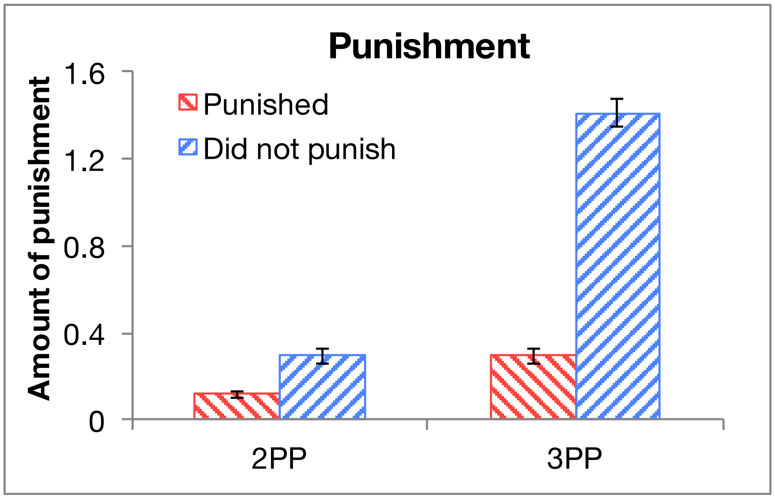

Figure 3. Results from Experiment 2. Participants read vignettes describing everyday situations in which a protagonist did or did not engage in punishment, either as a second or third party. Participants decided how much a potential punisher, either a second party or third party, should be punished for either engaging in punishment or not engaging in punishment. Punishment was assessed on a 0 to 9 scale, anchored at “No punishment at all” and “Extreme punishment”, respectively. Amount of punishment endorsed by participants in the second- or third-party condition by whether the protagonist punished or not. Error bars are SEM.

Overall, we again find a significant interaction between whether or not punishment occurred and whether the participant was assessing a second- or third-party (1000 sample bootstrap LRT $X^2(1) = 53.04$, $p = 0.001$, Odds Ratio = 5.60, 95% CI = 3.55 – 9.15) as well as a main effect of condition (1000 sample bootstrap LRT $X^2(1) = 51.23$, $p = 0.001$, Odds Ratio = 7.95, 95% CI = 5.09 – 12.9) and whether or not punishment occurred (1000 sample bootstrap LRT $X^2(1) = 54.56$, $p = 0.001$, Odds Ratio

Electronic copy available at: https://ssrn.com/abstract=3080990

= 8.87, 95% CI = 5.69 – 14.39). First, looking at the proportion of trials on which a participant endorsed HOP in the third-party condition, we find that participants were more likely to endorse punishment of those who do not engage in FOP than those who do engage in FOP (Did not punish: 52.2%, *SE* of proportion = 0.01; Did punish: 11.4%, *SEP* = 0.01; McNemar's $X^2(1, N = 1139) = 287.7$, $p < 0.001$, Odds Ratio = 8.23, 95% CI of OR = 6.45 – 10.51). We find a similar pattern when we look at the second-party condition (Did not punish: 11.3%, *SEP* = 0.01; Did punish: 5.5%, *SEP* = 0.01; McNemar's X^2 (1, $N = 1163$) = 22.9, $p < 0.001$, Odds Ratio = 1.67, 95% CI of OR = 1.35 – 2.06), albeit a weaker one.

Finally, participants were more likely to endorse HOP of third parties who failed to punish than second parties who failed to punish (McNemar's X^2 (1, $N = 1139$) = 292.7, $p < 0.001$, Odds Ratio = 8.44, 95% CI of OR = 6.6 – 10.78), and of third parties who did punish relative to second parties who did punish (McNemar's X^2 (1, $N = 1181$) = 24.6, $p < 0.001$, Odds Ratio = 1.7, 95% CI of OR = 1.38 – 2.09), albeit to a lesser extent. Thus, we find that whether or not FOP occurred influences whether people endorse HOP when they are a third-party, and when they are a second-party, though we find a much (4 times) larger effect for third parties; and that more people endorse HOP of third parties who fail to punish than second parties who fail to punish.

3.2.4 Endorsement of HOP among those endorsing HOP

Consistent with this difference in the proportion of individuals endorsing HOP between conditions, we find no difference between conditions in the amount of punishment endorsed given that one endorsed any punishment, either when FOP did occur (2PP: $M = 2.13$, $SEM = 0.22$; 3PP: $M = 2.58$, $SEM = 0.19$) or when FOP did not occur (2PP: *M* = 2.60, *SEM* = 0.18; 3PP: *M* = 2.71, *SEM* = 0.08; All Welch $t < 1.6$, all $p > 0.10$, all Cohen's $d < 0.26$).

3.2.5 Endorsement of HOP on the first trial

Because participants completed a series of trials, one possibility is that later responses were influenced by earlier trials. Of course, because participants read all vignettes in only one condition, this influence could not explicitly enhance differences between conditions. However, later trials could be influenced by reactions to prior vignettes. Looking at data from the first trial only provides a measure of participants' responding in the absence of such potential influence.

When analyzing the first trial, we continue to find a significant interaction between condition (2PP vs. 3PP) and whether or not the protagonist endorsed punishment (1000 sample bootstrap LRT $X^2(1) = 12.99$, $p < 0.001$, $\boldsymbol{\beta} = 0.13$, SE = 0.03, 95% CI = 0.06 – 0.19). We also find a significant main effect of condition (1000 sample bootstrap LRT $X^2(1) = 21.57$, $p < 0.001$, $\boldsymbol{\beta} = 0.16$, SE = 0.03, 95% CI = 0.10 – 0.23), and whether FOP occurred or not (1000 sample bootstrap LRT $X^2(1) = 15.63$, $p < 0.001$, $\boldsymbol{\beta} = 0.14$, SE = 0.03, 95% CI = 0.07 – 0.21). Looking at simple effects within each condition (2PP and 3PP), we find greater endorsement of HOP of non-punishing protagonists than of punishing protagonists in the 3PP condition (Did not punish: *M* = 0.78, *SEM* = 0.11; Did punish: *M* = 0.22, *SEM* = 0.07; Welch $t(330.27) = 4.39$, $p < 0.001$, Cohen's $d = 0.45$, 95% CI = 0.24 – 0.65) but not in the 2PP condition (Did not punish: *M* = 0.17, *SEM* = 0.06; Did punish: *M* = 0.14, *SEM* = 0.05; Welch $t(381.5) = 0.42$, $p = 0.68$, Cohen's $d = 0.04$, 95% CI = -0.16 – 0.24). Additionally, more HOP was endorsed when participants assessed third parties who failed to punish relative to participants who assessed second parties who failed to punish (Welch $t(288.16) = 5.10$, $p < 0.001$, Cohen's $d = 0.52$, 95% CI = 0.32 – 0.72) but not when the protagonist did punish (Welch $t(353.17) = 0.98$, $p = 0.33$, Cohen's $d = 0.1$, 95% CI = -0.1 – 0.3).

Finally, when looking only at the first trial in Experiment S2, we find that subjects endorse above-baseline HOP of third parties who fail to punish, but do not endorse above-baseline HOP of second parties who fail to punish. Thus, in the 3PP condition, we observe the same pattern of results as we found across all trails, but in the 2PP condition, the first trail fails to reveal above-baseline endorsement of HOP.

3.3 Exp 2. Discussion

Consistent with our Experiment 1 results, Experiment 2 provides evidence that participants endorse HOP, and also that HOP is endorsed more strongly in the context of 3PP than 2PP. Our key results are thus observed both in abstract incentive-compatible games and also in moral judgments of vignettes that are concrete and ordinary, but hypothetical.

In contrast to Experiment 1, we find some (minimal) endorsement of HOP in the context of 2PP when looking at punishment across trials, though not when examining the first trial only. We return to the question of the consistency of HOP in the context of 2PP following Experiment 3.

4. Experiment 3

Experiments 1 and 2 examine higher-order punishment (HOP) by focusing on prototypical cases: A harm occurs, someone has a chance to punish, and they either do or do not. However, this introduces a potential confound that complicates the interpretation of our results. Specifically, there is a difference in victimhood between our

Electronic copy available at: https://ssrn.com/abstract=3080990

second- and third-party conditions. In the second-party condition, the potential first-order punisher has *already been harmed*, either as a victim of theft (Exp. 1) or property damage (Exp. 2). Consequentially, participants might be less inclined to assign higher-order punishment to second parties who failed to punish, because doing so would impose further costs on this victim. In other words, participants asked to engage in HOP may reason, "This victim may be failing to uphold an important norm by declining to retaliate, but hasn't she already suffered enough?" In contrast, in the third-party condition, the potential first-order punisher was never harmed, and thus this concern would not apply.

To examine this possibility, we turn to a new type of situation in Experiment 3: Vignettes involving attempted, rather than completed, harms. Here, harm does not actually befall the victim (though all involved parties think it has at the time punishment can be enacted). Consequently, there is no difference between the second- and third-party conditions in whether the potential first-order punisher has been harmed.

For instance, in one of our vignettes John observes a large rock thrown by a stranger hit his car (2PP condition) or Steve's car (3PP condition). The rock has left what looks like a big scratch in the car's paint. John either does nothing in response (i.e. he does not engage in punishment) or he yells at the stranger and condemns their behavior. After the stranger has left, John inspects the paint and realizes that no damage has been done: the scratch is just a dirt mark. Thus, the potential first-order punisher in the 2PP condition (John, when John's car was hit) has been harmed no more than the potential first-order punisher in the 3PP condition (John, when Steve's car was hit).

Electronic copy available at: https://ssrn.com/abstract=3080990

4.1 Exp 3. Methods

Participants (*N* = 901) read through ten vignettes designed to closely match the structure of the economic game in Experiment 1 and the vignettes described in Experiment 2, except here vignettes involved an attempted but failed harm. The text of all vignettes was only minimally changed across the four conditions. As before, either the victim (second-party condition) or a third-party (third-party condition) is present immediately after the attempted harm takes place and either does (2PP *n* = 226; 3PP *n* = 226) or does not (2PP *n* = 226; 3PP *n* = 223) engage in punishment. The text of all cases can be found in the Supplemental Information. Participants were asked how much the potential first-order punisher should be punished using the same scale as Experiment 2. Participants were randomly assigned to one of the four conditions and read all vignettes for the experiment in that condition. After the final trial, participants completed attention check questions, for use in assessing data quality. Participants were excluded using the same criteria as Experiment 2. Based on these criteria, data from 163 participants (out of 901 18.1%; 2PP, did punish = 34; 2PP, did not punish = 41; 3PP, did punish = 38; 3PP, did not punish = 50) were discarded. Percentage excluded was similar across all four conditions (2PP, did punish: 15.0%; 2PP, did not punish: 18.1%; 3PP, did punish: 16.8%; 3PP, did not punish: 22.4%). Including all subjects does not change the overall pattern of results. All procedures were approved by the Harvard University Committee on the Use of Human Subjects.

Data analysis was carried out as specified in Experiment 2.

4.2 Exp 3. Results

4.2.1 Endorsement of HOP

Electronic copy available at: https://ssrn.com/abstract=3080990

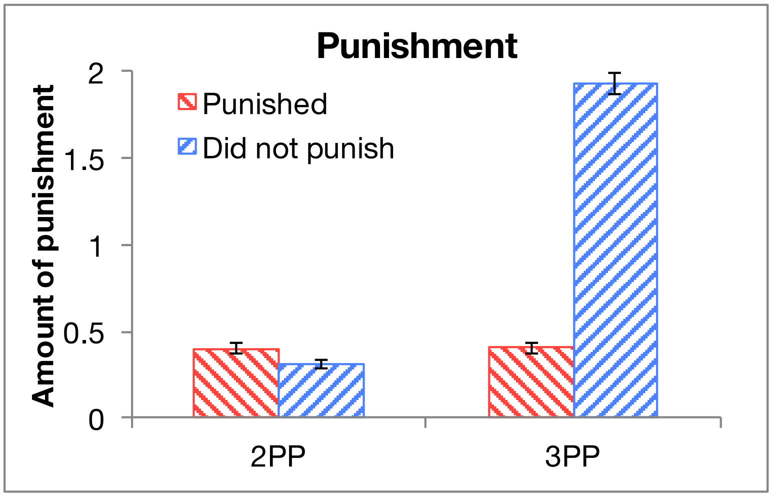

Figure 4. Results from Experiment 3. Participants read vignettes describing everyday situations in which a protagonist did or did not engage in punishment, either as a second or third party, in response to an attempted (but failed) harm. Participants decided how much a potential punisher, either a second party or third party, should be punished for either engaging in punishment or not engaging in punishment. Punishment was assessed on a 0 to 9 scale, anchored at “No punishment at all” and “Extreme punishment”, respectively. Amount of punishment endorsed by participants in the second- or third-party condition by whether the protagonist punished or not. Error bars are SEM.

We first analyze the amount of HOP endorsed using linear mixed-effects regression. We find a significant interaction between condition (2PP vs. 3PP) and whether or not the protagonist punished (1000 sample bootstrap LRT $X^2(1) = 74.78$, $p < 0.001$, **β** = 0.24, SE = 0.03, 95% CI = 0.18 – 0.29). We also find a significant main effect of condition (1000 sample bootstrap LRT $X^2(1) = 64.74$, $p < 0.001$, **β** = 0.23, SE = 0.03, 95% CI = 0.18 – 0.28), and whether FOP occurred or not (1000 sample bootstrap LRT $X^2(1) = 51.97$, $p < 0.001$, **β** = 0.21, SE = 0.03, 95% CI = 0.15 – 0.26). Given the presence of the interaction, we followed up the significant main effects with tests of the simple effects within each condition (2PP and 3PP).

Electronic copy available at: https://ssrn.com/abstract=3080990

First, looking at HOP in the 3PP condition, we observe greater endorsement of HOP of non-punishing protagonists than of punishing protagonists (Figure 3; Did not punish: *M* = 1.83, *SEM* = 0.15; Did punish: *M* = 0.47, *SEM* = 0.07; Welch *t*(243.15) = 8.46, *p* < 0.001, Cohen's *d* = 0.89, 95% CI = 0.67 – 1.11). We find the opposite pattern in the 2PP condition (Did not punish: *M* = 0.29, *SEM* = 0.07; Did punish: *M* = 0.49, *SEM* = 0.06; Welch *t*(371.65) = 2.1, *p* = 0.034, Cohen's *d* = 0.22, 95% CI = 0.02 – 0.42). Finally, greater HOP was endorsed when participants assessed third parties relative to participants who assessed second parties, but only when they failed to punish (Welch *t*(246.9) = 9.58, *p* < 0.001, Cohen's *d* = 1.01, 95% CI = 0.79 – 1.23) and not when they did punish (Welch *t*(376.74) = 0.14, *p* = 0.89, Cohen's *d* = 0.01, 95% CI = -0.19 – 0.22).

4.2.2 Endorsement of HOP relative to baseline

Like in the case of Experiment 2, in a separate study, we also examined the extent to which subjects endorsed HOP of non-punishers in the context of 3PP and 2PP, relative to a neutral baseline condition (in which subjects evaluated an entirely uninvolved individual). We found greater endorsement of HOP of non-punishers in the 3PP condition, but not in the 2PP condition, relative to entirely uninvolved individuals (in the neutral baseline condition) (see Supplemental Experiment S3). Thus, like in Experiment S2, Experiment S3 find above-baseline endorsement of HOP of third parties who do not punish selfishness. However, unlikely in Experiment S2, Experiment S3 does not find above-baseline endorsement of HOP of second parties who not punish.

4.2.3 Proportion of trials on which HOP was endorsed

We again also analyze the proportion of trials on which participants endorsed HOP, as in Experiment 2. Overall, we again find a significant interaction between whether or not punishment occurred and whether the participant was assessing a second or third-party (1000 sample bootstrap LRT $X^2(1) = 108.52$, $p = 0.001$, Odds Ratio = 4.01, 95% CI = 3.06 – 5.37) as well as a main effect of condition (1000 sample bootstrap LRT $X^2(1) = 92.75$, $p = 0.001$, Odds Ratio = 3.84, 95% CI = 2.94 – 5.14) and whether or not punishment occurred (1000 sample bootstrap LRT $X^2(1) = 18.13$, $p = 0.001$, Odds Ratio = 1.60, 95% CI = 1.23 – 2.08). First, looking at the proportion of trials on which a participant endorsed HOP in the third-party condition, we find that participants were more likely to endorse HOP of those who failed to engage in FOP than those who did engage in FOP (Did not punish: 54.5%, *SE* of proportion = 0.01; Did punish: 15.3%, *SEP* = 0.01; McNemar's $X^2(1, N = 1325) = 288.7$, $p < 0.001$, Odds Ratio = 6.78, 95% CI of OR = 5.43 – 8.46). We find the opposite pattern when we look at the second-party condition (Did not punish: 11.4%, *SEP* = 0.01; Did punish: 15.8%, *SEP* = 0.01; McNemar's X^2 (1, $N = 1452$) = 21.76, $p < 0.001$, Odds Ratio = 1.56, 95% CI of OR = 1.3 – 1.89). Finally, participants were more likely to endorse HOP of third parties who failed to punish than second parties who failed to punish (McNemar's X^2 (1, $N = 1325$) = 257.15, $p < 0.001$, Odds Ratio = 5.93, 95% CI of OR = 4.77 – 7.37), and of third parties who did punish relative to second parties who did punish (McNemar's X^2 (1, $N = 1336$) = 12.77, $p < 0.001$, Odds Ratio = 1.43, 95% CI of OR = 1.17 – 1.74). Thus, we find that whether or not FOP occurs influences how many people endorse HOP when they are a third-party and when they are a second-party, but in opposite ways; and that more people endorse HOP of third parties who fail to punish than second parties who fail to punish.

4.2.4 Endorsement of HOP among those endorsing HOP

Electronic copy available at: https://ssrn.com/abstract=3080990

In spite of this difference in the proportion of individuals endorsing HOP between conditions, we continue to find, given that any HOP was endorsed, greater HOP was endorsed in the context of 3PP. In the 3PP condition, we observe greater endorsement of HOP of non-punishing protagonists than of punishing protagonists (Did not punish: *M* = 3.51, *SEM* = 0.07; Did punish: *M* = 2.55, *SEM* = 0.12; Welch *t*(548.46) = 6.99, *p* < 0.001, Cohen's *d* = 0.46, 95% CI = 0.33 – 0.60). We find no difference in the 2PP condition (Did not punish: *M* = 2.76, *SEM* = 0.15; Did punish: *M* = 2.53, *SEM* = 0.11; Welch *t*(395) = 1.23, *p* = 0.22, Cohen's *d* = 0.11, 95% CI = -0.07 – 0.29). We also continue to find greater endorsement of HOP of third parties who fail to punish than second parties who fail to punish (Welch *t*(308.34) = 4.39, *p* < 0.001, Cohen's *d* = 0.33, 95% CI = 0.18 – 0.49), but no difference for protagonists who did punish (Welch *t*(594.97) = 0.15, *p* = 0.88, Cohen's *d* = 0.01, 95% CI = -0.15 – 0.17).

4.2.5 Endorsement of HOP on the first trial

We also again examine responses to the first trial only. We continue to find a significant interaction between condition (2PP vs. 3PP) and whether or not the protagonist punished (1000 sample bootstrap LRT $X^2(1) = 28.85$, *p* < 0.001, **β** = 0.19, SE = 0.03, 95% CI = 0.12 – 0.25). We also find a significant main effect of condition (1000 sample bootstrap LRT $X^2(1) = 18.71$, *p* < 0.001, **β** = 0.15, SE = 0.03, 95% CI = 0.08 – 0.22), and whether FOP occurred or not (1000 sample bootstrap LRT $X^2(1) = 8.83$, *p* = 0.004, **β** = 0.10, SE = 0.03, 95% CI = 0.04 – 0.17). Looking at simple effects within each condition (2PP and 3PP), we find greater endorsement of HOP of non-punishing protagonists than of punishing protagonists in the 3PP condition (Did not punish: *M* = 1.49, *SEM* = 0.17; Did punish: *M* = 0.51, *SEM* = 0.10; Welch *t*(288.34) =

4.98, *p* < 0.001, Cohen's *d* = 0.52, 95% CI = 0.31 – 0.74) but not in the 2PP condition (Did not punish: *M* = 0.32, *SEM* = 0.09; Did punish: *M* = 0.57, *SEM* = 0.11; Welch *t*(363.76) = 1.75, *p* = 0.08, Cohen's *d* = 0.18, 95% CI = -0.02 – 0.38). And, greater HOP was endorsed when participants assessed third parties who failed to punish than when participants assessed second parties who failed to punish (Welch *t*(265.65) = 6.15, *p* < 0.001, Cohen's *d* = 0.65, 95% CI = 0.44 – 0.86) but not when the protagonist did punish (Welch *t*(376.75) = 0.38, *p* = 0.70, Cohen's *d* = 0.04, 95% CI = -0.16 – 0.24). Thus, we observe the same pattern of results in the 3PP condition, but now no difference in the 2PP condition. Importantly, we continue to find an overall interaction between condition and whether the protagonist punished or not. These results are mirrored in Experiment S3, where first trial data reveals above-baseline HOP of third parties who fail to punish, but not second parties who fail to punish.

4.3 Exp 3. Discussion

Experiment 3 suggests that HOP is not stronger in the context of 3PP than 2PP simply because of the outcome asymmetry between observers and victims. Vignettes involving attempted but failed harms eliminate this asymmetry, yet we replicate (with a similar effect size) the finding from Experiments 1 and 2 that participants endorse HOP of non-punishing observers more than they endorse HOP of non-punishing victims. In Experiment 3, regarding endorsement of HOP in the context of 2PP, we find results that are inconsistently significant. However, when we do find significant effects, they show the opposite pattern compared with Experiments 1 and 2: Greater endorsement of HOP of those who do punish than those who do not. In summary, then, we consistently observe an HOP effect for third parties and only inconsistently for 2PP, and we find that

Electronic copy available at: https://ssrn.com/abstract=3080990

more HOP is assigned to third parties who fail to punish than second parties who fail to punish**.** We discuss the overall pattern of results regarding HOP of 2PP below.

5. Experiment 4

In Experiment 4 we explore the range of circumstances under which subjects selectively endorse HOP of third parties who fail to punish. In particular, we investigate the role of two factors: The type of violation involved and the relationship between the third-party and the victim of harm. In Experiments 2 and 3, although we used violations of an everyday nature, they were severe enough to involve property damage or potential physical harm, and were criminal violations. Thus, it is unclear whether our results would generalize to the context of less severe violations. In Experiment 4, we investigate HOP in the context of less severe, non-criminal violations (e.g. offensive language, taking someone's lunch out of a shared refrigerator, or invasion of privacy).

Additionally, Experiment 4 investigates whether, in the context of 3PP, the relationship between the third-party and the victim influences endorsement of HOP. Whereas in Experiment 1 the victim and third-party were strangers, the victim was either a friend, co-worker, neighbor, or classmate in Experiments 2 and 3. Thus, our vignettes have described cases in which there both is and is not a relationship between the third-party and the victim, but we have not yet systematically manipulated this factor. Yet there is reason to believe the relationship between the third-party and the victim might be important.

Outside of formal institutions, we might expect that people see third-party punishment as more injunctively normative when the third-party observer has a close relationship with the victim. Additionally, in contemporary Western societies, it is

Electronic copy available at: https://ssrn.com/abstract=3080990

common for social organizations (e.g., clubs, corporations or governments) to institutionalize punishment by third parties in dedicated roles (e.g., chairpersons, compliance officers, police and judges) (Cushman, 2015). And punishment from third parties in these roles may also be seen as injunctively normative: Indeed, an individual in one of these roles who fails to perform their duty can expect to lose their title, office, job or salary (Hilbe, Traulsen, Rohl, & Milinski, 2014), and face intense condemnation in the community. Thus, when a third party fails to punish, we might expect more HOP in contexts where there is a personal relationship between the victim and the third party, or where the third party is an authority position. To test this hypothesis, in Experiment 4 we include three third-party conditions, in which the third party is (i) a stranger to the victim, (ii) the victim's friend, or (iii) an authority figure (e.g. the victim's boss, or the manager of the location at which the violation towards the victim takes place).

5.1 Exp 4. Methods

Participants (*N* = 1804) read through eight vignettes designed to be similar to the vignettes used in Experiments 2 and 3 with a few differences. We manipulated two main factors. First, as in our other experiments, we manipulate whether a first-order punisher either does or does not engage in punishment. Second, we manipulate the relationship between the first-order punisher and the victim, with four levels of relationship. The first-order punisher is either (1) the victim themselves (*n* = 457), (2) a third-party stranger (*n* = 444), (3) a third-party friend (*n* = 461) or (4) a third-party authority figure (e.g. a boss) (*n* = 449). In all cases, the first-order punisher is present immediately after the harm takes place and either does or does not engage in punishment. The text of all cases can be found in the Supplemental Information.

Electronic copy available at: https://ssrn.com/abstract=3080990

Participants were asked how much the potential first-order punisher should be punished using the same scale as Experiments 2 and 3. Participants were randomly assigned to one of the eight conditions and read all vignettes for the experiment in that condition. After the final trial, participants completed attention check questions, for use in assessing data quality. Participants were excluded using the same criteria as for Experiments 2 and 3. Based on these criteria, data from 261 participants (out of 1804 14.4%) were discarded. Percentage excluded was similar across all eight conditions (2PP, did punish: *n* = 45, 19.7%; 2PP, did not punish: *n* = 38, 16.6%; 3PP stranger, did punish: *n* = 28, 12.7%; 3PP stranger, did not punish: *n* = 33, 14.8%; 3PP friend, did punish: *n* = 30, 13.6%; 3PP friend, did not punish: *n* = 26, 11.1%; 3PP authority, did punish: *n* = 34, 15.0%; 3PP authority, did not punish: *n* = 27, 12.2%). Including all subjects does not change the overall pattern of results. All procedures were approved by the Harvard University Committee on the Use of Human Subjects.

Data analysis was carried out as specified in Experiment 2 and 3, with a few exceptions. First, we begin by running models comparing the 2PP cases to all 3PP cases (ignoring the relationship of the third party to the victim). We next compare the 2PP cases against the 3PP stranger cases alone. Finally, we compare the 3PP cases against each other, comparing 3PP friend with 3PP stranger and 3PP friend with 3PP boss.

Sample size, exclusionary criteria and analysis approach were pre-registered at http://aspredicted.org/blind.php?x=2uc7yw

Electronic copy available at: https://ssrn.com/abstract=3080990

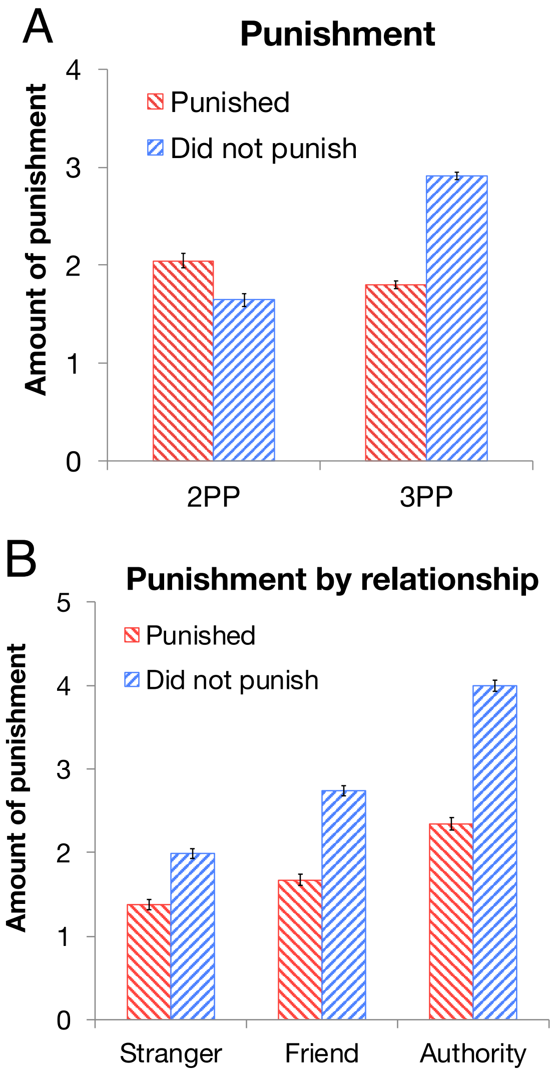

Figure 5. Results from Experiment 4. Participants read vignettes describing everyday situations in which a protagonist did or did not engage in punishment, either as a second or third party, in response to a non-criminal violation. Participants decided how much a potential punisher, either a second party or third party, should be punished for either engaging in punishment or not engaging in punishment. Punishment was assessed on a 0 to 9 scale, anchored at “No punishment at all” and “Extreme punishment”, respectively. **A:** Amount of punishment endorsed by participants in the second-party condition versus the (combined) third-party conditions by whether the protagonist punished or not. **B:** Amount of punishment endorsed within the third-party conditions by the relationship between the third party and the victim and whether or not the third party punished. Error bars are SEM.

5.2 Exp 4. Results

Electronic copy available at: https://ssrn.com/abstract=3080990

5.2.1 Comparing 2PP against 3PP overall

5.2.1.1 Endorsement of HOP

We begin by comparing the 2PP cases against all 3PP cases, first analyzing the amount of HOP endorsed using linear mixed-effects regression. We find a significant interaction between condition (2PP vs. 3PP) and whether or not the protagonist punished (1000 sample bootstrap LRT $X^2(1) = 31.98$, $p < 0.001$, $\boldsymbol{\beta} = 0.12$, SE = 0.02, 95% CI = 0.8 – 0.16). We also find a significant main effect of condition (1000 sample bootstrap LRT $X^2(1) = 41.07$, $p < 0.001$, $\boldsymbol{\beta} = 0.08$, SE = 0.02, 95% CI = 0.04 – 0.12), and whether FOP occurred or not (1000 sample bootstrap LRT $X^2(1) = 15.15$, $p < 0.001$, $\boldsymbol{\beta} = 0.14$, SE = 0.02, 95% CI = 0.10 – 0.18). Given the presence of the interaction, we followed up the significant main effects with tests of the simple effects within each condition (2PP and 3PP).

First, looking at HOP across the 3PP conditions, we observe greater endorsement of HOP of non-punishing protagonists than of punishing protagonists (Did not punish: $M = 2.91$, $SEM = 0.04$; Did punish: $M = 1.80$, $SEM = 0.04$; Welch $t(1128.7) = 8.57$, $p < 0.001$, Cohen's $d = 0.50$, 95% CI = 0.38 – 0.62). We find no significant difference between punishing and non-punishing protagonists in the 2PP condition (Did not punish: $M = 1.64$, $SEM = 0.07$; Did punish: $M = 2.05$, $SEM = 0.07$; Welch $t(360.79) = 1.66$, $p = 0.10$, Cohen's $d = 0.17$, 95% CI = -0.03 – 0.38). We observe the same pattern of results when looking only at first trial response (see Supplemental Results). Thus, we replicate our key pattern of results from Experiments 1-3.

While this analysis collapses across our three 3PP conditions, as a more conservative test, we also conduct a secondary analysis in which we compare our 2PP

Electronic copy available at: https://ssrn.com/abstract=3080990

condition to our 3PP “stranger” condition (in which we would expect 3PP to be the least injunctively normative). In this analysis, we continue to find an interaction between condition and whether FOP occurred or not, both when looking at the first trial only and all trials, and when analyzing the proportion of trials on which participants endorsed HOP (see Supplemental Results). Thus, even when the third party is a stranger to the victim, we observe the same pattern whereby subjects selectively endorse HOP of third parties who fail to punish.

Finally, we note that contrary to our prior experiments, in Experiment 4 we find that among those endorsing HOP, more HOP was endorsed in the context of 2PP, and greater HOP was endorsed in cases where FOP did not occur (see Supplemental Results).

5.2.1.2 Proportion of trials on which HOP was endorsed

We again also analyze the proportion of trials on which participants endorsed HOP, as in Experiments 2 and 3. Overall, we again find a significant interaction between whether or not punishment occurred and whether the participant was assessing a second or third party (1000 sample bootstrap LRT $X^2(1) = 72.09$, $p < 0.005$, Odds Ratio = 3.15, 95% CI = 2.40 – 4.16) as well as a main effect of condition (1000 sample bootstrap LRT $X^2(1) = 46.71$, $p < 0.005$, Odds Ratio = 2.40, 95% CI = 1.84 – 3.15) and whether or not punishment occurred (1000 sample bootstrap LRT $X^2(1) = 173.57$, $p < 0.005$, Odds Ratio = 6.1, 95% CI = 4.59 – 8.80). First, looking at the proportion of trials on which a participant endorsed HOP across the third party conditions, we find that participants were more likely to endorse HOP of those who failed to engage in FOP than those who did engage in FOP (Did not punish: 75.2%, *SE* of proportion = 0.01; Did punish: 38.9%, *SEP* = 0.01; McNemar's X^2(1, *N* = 9352) = 585.27, *p* < 0.001, Odds Ratio = 3.90, 95% CI of OR = 3.49 – 4.35). We find no significant difference in the second-party condition and, if anything, directionally observe the opposite pattern (Did not punish: 39.3%, *SEP* = 0.01; Did punish: 43.1%, *SEP* = 0.01; McNemar's X^2 (1, *N* = 2992) = 0.68, *p* = 0.41, Odds Ratio = 1.08, 95% CI of OR = 0.90 – 1.30). Thus, we find that whether or not FOP occurs influences whether HOP is endorsed only when the punisher is a third party and not when they are a second party.

In total, these results support and extend the results of Experiments 2 and 3 by generalizing them to contexts where the original violation is less severe, and non-criminal in nature. In the final sections on our Experiment 4 results, we compare our different 3PP conditions in order to investigate the role of the relationship between the third party and the victim.

5.2.2 Comparing third-party strangers with third-party friends

We next investigate the role that the relationship between the third party and the victim plays in endorsement of HOP. First, for cases in which the third party is a stranger to the victim compared to those in which the third party is a friend of the victim, we do not find a significant interaction between relationship and whether or not the protagonist punished (1000 sample bootstrap LRT X^2(1) = 2.37, *p* = 0.13, $\boldsymbol{\beta}$ = 0.05, SE = 0.03, 95% CI = -0.01 – 0.10) (although we do observe a directionally larger effect of whether FOP occurred within the "friend" condition, as might be expected). We do find a significant main effect of relationship (1000 sample bootstrap LRT X^2(1) = 31.91, *p* < 0.001, $\boldsymbol{\beta}$ = 0.11, SE = 0.03, 95% CI = 0.05 – 0.16), and whether FOP occurred (1000 sample bootstrap LRT X^2(1) = 12.74, *p* < 0.001, $\boldsymbol{\beta}$ = 0.17, SE = 0.03, 95% CI = 0.11 –

Electronic copy available at: https://ssrn.com/abstract=3080990

0.23). Looking at HOP when the third party is the victim's friend, we observe greater endorsement of HOP of non-punishing protagonists than of punishing protagonists (Did not punish: *M* = 2.74, *SEM* = 0.06; Did punish: *M* = 1.67, *SEM* = 0.07; Welch *t*(377.01) = 4.95, *p* < 0.001, Cohen's *d* = 0.50, 95% CI = 0.30 – 0.70). We find the same pattern when the third party is a stranger (Did not punish: *M* = 1.99, *SEM* = 0.06; Did punish: *M* = 1.38, *SEM* = 0.06; Welch *t*(360.36) = 3.01, *p* < 0.005, Cohen's *d* = 0.31, 95% CI = 0.11 – 0.51). Finally, greater HOP was endorsed when a friend failed to punish than when a stranger failed to punish (Welch *t*(395.92) = 3.97, *p* < 0.001, Cohen's *d* = 0.40, 95% CI = 0.20 – 0.60) but we find no significant difference between the cases where a friend versus stranger chose to punish (Welch *t*(379.19) = 1.31, *p* = 0.19, Cohen's *d* = 0.13, 95% CI = -0.07 – 0.34). We find the same pattern of results when examining first trial responses only as well as when looking at the proportion of trials on which HOP was endorsed (see Supplemental Results).

5.2.3 Comparing third-party strangers with third-party authorities

Next, we compare cases in which the third party is an authority figure in the relevant context against cases in which the third party is a stranger to the victim. We find a significant interaction between whether the third-party protagonist was a stranger or authority and whether or not the protagonist punished (1000 sample bootstrap LRT $X^2(1) = 11.66$, *p* < 0.001, **β** = 0.09, SE = 0.03, 95% CI = 0.04 – 0.15). We also find a significant main effect of whether the third party was a stranger or authority (1000 sample bootstrap LRT $X^2(1) = 89.08$, *p* < 0.001, **β** = 0.27, SE = 0.03, 95% CI = 0.22 – 0.32), and whether FOP occurred or not (1000 sample bootstrap LRT $X^2(1) = 53.11$, *p* < 0.001, **β** = 0.21, SE = 0.03, 95% CI = 0.15 – 0.26). Looking at HOP when the 3rd party

Electronic copy available at: https://ssrn.com/abstract=3080990

is an authority, we observe greater endorsement of HOP of non-punishing protagonists than of punishing protagonists (Did not punish: M = 4.00, SEM = 0.07; Did punish: M = 2.34, SEM = 0.07; Welch $t(342.24) = 7.28$, $p < 0.001$, Cohen's d = 0.74, 95% CI = 0.53 – 0.95). We find the same pattern when the 3rd party is a stranger (Did not punish: M = 1.99, SEM = 0.06; Did punish: M = 1.38, SEM = 0.06; Welch $t(360.36) = 3.01$, $p < 0.005$, Cohen's d = 0.31, 95% CI = 0.11 – 0.51). Finally, greater HOP was endorsed for third-party authorities than third-party strangers both when they failed to punish (Welch $t(382.86) = 10.96$, $p < 0.001$, Cohen's d = 1.12, 95% CI = 0.90 – 1.33) and when they did punish (Welch $t(371.84) = 3.98$, $p < 0.001$, Cohen's d = 0.41, 95% CI = 0.20 – 0.61). We find the same pattern of results when analyzing first trial responses only (see Supplemental Results). When analyzing the proportion of trials on which participants endorsed HOP, we do not find an interaction between relationship type and whether the protagonist punished (p = 0.13), although the qualitative pattern is the same (see Supplemental Results).

5.3 Exp 4. Discussion

Experiment 4 builds upon and extends the results of Experiment 2 and 3 in important ways. First, we replicate our main finding, that HOP is selectively directed at the failure to engage in 3PP, in the context of less severe and non-criminal violations. This extends the signature of selective HOP from economic games and both intentional and attempted cases of property damage and physical harm to cases of non-criminal norm violations. Additionally, within the context of third-party punishment, Experiment 4 explores the influence of the relationship between the third party and the victim on endorsement of HOP. While we observe that HOP is selectively targeted at non-punishers in all three 3PP conditions, this pattern is directionally strongest when an authority fails to punish, weaker when a friend of the victim fails to punish, and weaker still when a stranger fails to punish. And, this pattern is significantly stronger in the authority condition than the stranger condition. We discuss the implications of this pattern more fully below**.**

6. General Discussion

Humans punish a wide variety of antisocial behaviors (Balafoutas, Nikiforakis, & Rockenbach, 2014; Balliet & Van Lange, 2013; Bone & Raihani, 2015; Buckholtz et al., 2008; Carpenter & Matthews, 2009; Cushman et al., 2009; de Quervain et al., 2004; Fehr & Fischbacher, 2004b; Fehr & Gächter, 2002; FeldmanHall et al., 2014; Gächter et al., 2008; Henrich et al., 2010, 2006; Martin & Cushman, 2015, 2016b; Mathew & Boyd, 2011; McCullough et al., 2013; Mendoza et al., 2014; Morris et al., 2017; Treadway et al., 2014). Here, we have investigated the question of when such punishment is seen as injunctively normative, and when it is not.

We approached this question by investigating when non-punishers are, themselves, punished. Presumably the existence of such “higher-order punishment” (HOP) indicates that punishment is seen as injunctively normative. We find evidence that such higher-order punishment does occur: Individuals both endorse higher-order punishment in hypothetical vignettes and enact HOP in incentivized economic games. Critically, however, we do not observe equal HOP in all contexts. Rather, our results indicate that HOP is applied to third-party observers who fail to sanction harm-doers relatively frequently, but is less frequently applied to third-party observers who do sanction harm-doers. Additionally, HOP is diminished and inconsistent in the context of

Electronic copy available at: https://ssrn.com/abstract=3080990

second-party punishment. This pattern suggests that some people see third-party punishment as injunctively normative, and that third-party punishment is seen as substantially more injunctively normative than second-party punishment.

Specifically, regarding HOP in the context of 2PP, in Experiments 1 and 4 (and Experiment S1) we found no selective HOP of non-punishing second parties. In Experiment 2 (and Experiment S2), we found some HOP of non-punishing second parties, though only when looking at responses across all trials and not when looking at data from the first trial only. In Experiment 3 we found HOP of those who *do* engage in 2PP (i.e. greater punishment of those punishing as a harmed victim), but only when looking at responses overall and not when looking at first trial responses. And, we found no HOP of non-punishing second parties in Experiment S3.

5.1 Relation to prior research

Prior HOP research has focused on the relationship between higher-order punishment and the provisioning of public goods, especially using the public goods game (Cinyabuguma et al., 2006; Fu et al., 2017; Kiyonari & Barclay, 2008). As noted above, these studies failed to find evidence of HOP of non-punishers (Cinyabuguma et al., 2006; Fu et al., 2017; Kiyonari & Barclay, 2008). Critically, participants in the public goods game serve as a mix between a second- and third-party, in that defectors harm each individual (thus making them a victim) but also harm all other individuals (making them a third-party). Thus, our results do not have strong implications for the prevalence of HOP in the public goods game. We note, however, that the public goods game may be useful in adjudicating between two possible interpretations for our results.

Electronic copy available at: https://ssrn.com/abstract=3080990

Specifically, there may be a norm for *punishment of violations with victims that include others* (i.e., including victims other than oneself), which would encompass PGGs, or there may be a norm for *punishment of any violations with victims not including oneself*, which would not encompass PGGs. The lack of evidence for HOP in the context of the PGG suggests that the latter possibility is more likely, though a more direct comparison may be useful.

Our results also relate to the broader literature on perceptions of and reactions to third-party punishers, which has found mixed evidence across contexts. A field experiment using littering as the violation found that those who engaged in 3PP and were subsequently in need were helped at no greater rate than a random individual in need (Balafoutas et al., 2014). In contrast, third-parties who punish unfairness or theft in the context of an economic game are rewarded more than third parties who do nothing (though third parties who helped are rewarded most) (Raihani & Bshary, 2015b) and are viewed more positively than those who do nothing (though less positively than those who compensate the victim) (Patil, Dhaliwal, & Cushman, 2018). These results suggest that responses to third-party punishment may vary across contexts, but also support the idea that third-party punishers can be viewed positively. Our work builds on this research by measuring higher-order punishment, which we see as an especially unambiguous index that first-order punishment is seen as normative.

Finally, prior work has demonstrated that individuals view third-parties as having a “role” or duty to punish in some contexts (Eriksson, Strimling, & Ehn, 2013; Strimling & Eriksson, 2014). In Experiment 4, we explicitly test this possibility in the context of vignettes describing everyday non-criminal violations. We find that HOP is selectively

Electronic copy available at: https://ssrn.com/abstract=3080990

targeted at non-punishers in all three 3PP conditions, but this pattern is directionally strongest when the third party is an authority such as a boss, intermediate when the third party is a friend, and weakest when the third party is a stranger (and this pattern is significantly stronger in the authority condition than the stranger condition).

These results suggest that the norm for third-party punishment is stronger in the context of certain duties or roles. In our vignettes, having a position of authority is seen as conferring a heightened obligation to punish. Nonetheless, we observe HOP targeted at third parties who fail to punish even when the third party is a stranger to the victim, demonstrating that such a relationship is not necessary for at least some people to view punishment as normative.

5.2 The ultimate function of HOP and its selective application

Why should people see punishment as normative, and engage in HOP of non-punishers? And, more specifically, why should people specifically see 3PP as normative, and engage in HOP of third parties who fail to punish? Our studies do not directly speak to the function of HOP, or provide a functional explanation for why HOP is selectively applied to 3PP. Nevertheless, our results do hold some circumstantial clues. Here we describe several potential functions of HOP, and we consider the extent to which each of these would predict relatively selective HOP (following from a relatively selective punishment norm) that applies to 3PP, but is mostly absent in cases of 2PP.

We begin this discussion by observing that HOP is, in fact, a special variety of 3PP. If somebody punishes an individual who themselves failed to punish (e.g., a social club kicks out a member who, as a boss, failed to police sexual harassment within a separate organization), then the “higher-order punisher” (e.g., the club members) is not punishing to avenge a harm to *themselves*, but rather a harm to some *third party* (or, more broadly, a norm violation). When we ask the question “why do people engage in HOP?” (e.g. “Why does the club care? Why should it punish the boss?”), we are in fact asking a specific variety of the question “why do people engage in 3PP”—that is, why do people ever punish when they are not, themselves, the direct victims of a harmful act? Thus, a discussion of mechanisms for 3PP may be informative for considering why (and when) our subjects enacted HOP.

Several mechanisms have been proposed to explain 3PP. Third-party punishers may benefit via (i) signaling prosociality and trustworthiness (Barclay, 2006; Gordon, Madden, & Lea, 2014; Horita, 2010; Nelissen, 2008), including through costly signaling (Brandt et al., 2006; Jordan, Hoffman, Bloom, & Rand, 2016; Jordan & Rand, 2017; Raihani & Bshary, 2015b, 2015a), (ii) signaling a willingness to retaliate when harmed directly (Krasnow, Cosmides, Pedersen, & Tooby, 2012; Krasnow, Delton, Cosmides, & Tooby, 2016), (iii) earning higher-order rewards or avoiding higher-order punishments, and (iv) receiving direct reciprocity from the victims of the punished transgressions (Jordan & Rand, 2017). It has also been proposed that 3PP might evolve via cultural group selection (because 3PP deters wrongdoing and promotes social welfare) despite being strictly costly to the individual (Fehr, Fischbacher, & Gächter, 2002; Gintis, Bowles, Boyd, & Fehr, 2003).

Thus, insofar as HOP is a special case of 3PP, HOP could be motivated by any of the above mechanisms. But why would any of these mechanisms result in the observed selectivity of HOP—an incentive to punish failures of 3PP, but not failures of 2PP?

In order to address this question, it is helpful to note that 2PP is intrinsically incentivized by a very powerful and unique benefit: it serves to deter future transgressions towards the punisher (Clutton-Brock & Parker, 1995; Raihani, Thornton, & Bshary, 2012; Trivers, 1971). In contrast, however, the same is not intrinsically true of 3PP, so 3PP often requires some other incentive. Put simply, sticking up for yourself is more likely to prevent people from harming you than is sticking up for others.

To illustrate this point, consider, again, an example. If Janet steals from Tom and then gets punished by him, deterrence theory suggests that Janet will be less likely to steal from Tom in the future. This, of course, is a direct and unique benefit to Tom—a special benefit enjoyed by those who stand up for themselves, engaging in 2PP. What about 3PP—does it confer any analogous benefit? If Lisa punishes Janet for stealing from Tom, she too might benefit from the deterrence-related consequences of her punishment. By punishing Janet, Lisa might generally deter stealing in society (benefiting everyone including Lisa), or even specifically deter stealing against Lisa (as per the aforementioned mechanism in which 3PP serves as a signal of one's willingness to enact 2PP). Yet, these benefits are less unique to Lisa and, we presume, relatively less certain.

In sum, people face strong and direct personal deterrence-based incentives for second-party punishment, while deterrence-based benefits apply less reliably to third-party punishment. For this reason, we might think of 2PP as a relatively more self-oriented action, while 3PP is relatively more other-oriented. And if 3PP is other-oriented

Electronic copy available at: https://ssrn.com/abstract=3080990

while 2PP is not, we can think of declining to punish as a *selfish* choice in the context of 3PP but not 2PP.

The fact that 2PP can be thought of as a basically self-interested action, while 3PP can be thought of as a more prosocial action, may explain why 3PP is uniquely enforced by some degree of HOP. Above, we argued that HOP is a special case of 3PP, and thus that any of the generic mechanisms that have been proposed to explain 3PP may also explain HOP. However, these previously-proposed explanations for 3PP do not explain indiscriminate 3PP directed at any behavior; rather, they specifically explain 3PP directed at failures to act prosocially. In particular, each of these explanations rely on the idea that 3PP is socially valuable because it promotes prosocial behavior and deters selfish behavior.

First, 3PP has been proposed to signal prosociality and trustworthiness by signaling an individual’s incentives for prosocial behavior (Jordan et al., 2016; Jordan & Rand, 2017), but only 3PP that is directed at selfishness (i.e., enforces prosociality) should signal prosocial incentives. Second, 3PP has been proposed to signal a willingness to retaliate and thus deter the punished behavior (Krasnow et al., 2012, 2016), but this is only beneficial when the 3PP is directed at harmful behavior (i.e., selfishness). Third, 3PP has been proposed to confer higher-order rewards or prevent higher-order punishments. However, there are an infinite number of possible social norms regarding higher-order reward and punishment, and it has been proposed that equilibrium selection mechanisms specifically select for norms that promote prosocial behavior and deter selfishness (Henrich, 2006). Fourth, 3PP has proposed to elicit direct reciprocity from victims of punished transgressions—but this concept requires

Electronic copy available at: https://ssrn.com/abstract=3080990

punishment to be directed at selfish behavior that creates a victim. Finally, 3PP has been proposed to result from cultural group selection, which should specifically favor 3PP that helps the group by promoting prosocial behavior and deterring selfishness (Fehr et al., 2002; Gintis et al., 2003). Of course, it is worth noting that not all punishment by third-parties delivers social benefits in this way: individuals can also engage in “antisocial” punishment, or punishment of altruistic acts (Brañas-Garza, Espín, Exadaktylos, & Herrmann, 2014; Herrmann, Thöni, & Gächter, 2008). Nevertheless, because 3PP is frequently targeted preferentially at selfish or immoral behavior, explanations for 3PP have tended to focus on the social benefits provided.

In sum, current explanations for 3PP depend on the fact that 3PP delivers social benefits. Insofar as HOP is a special case of 3PP, we should specifically expect HOP to enforce “prosocial” acts of punishment. This implies HOP will be selectively applied to *third parties* who fail to punish, because 3PP is less likely to confer personal deterrence-based benefits and thus is more other-oriented, whereas 2PP tends to confer more personal benefits and is thus more selfish.

Notably, our analysis predicts that people should be sufficiently motivated to engage in 2PP by direct and personal benefits, and in the absence of social benefits such as negative or positive reciprocity (Bone & Raihani, 2015; McCullough et al., 2013). Consistent with this prediction, second parties are more tenacious punishers—seeking to punish when they could easily avoid it—whereas third parties are more likely to shirk on punishment if they are able to get away with it (Kriss, Weber, & Xiao, 2015). It also predicts that 3PP should be especially sensitive to the possibility of social benefits. Consistent with this prediction, third-party parties seem to be particularly

Electronic copy available at: https://ssrn.com/abstract=3080990

motivated by whether others will know about their punishment (Kurzban, Descioli, & Obrien, 2007).

Although we propose that HOP may help to stabilize 3PP, we certainly do not propose that it is the only such stabilizing force, or that it is sufficient. To the contrary, we expect that HOP promotes 3PP in conjunction with other mechanisms, several of which we described above. Consistent with this possibility, our experiments elicited relatively low absolute amounts of HOP, which may suggest that HOP is not an especially strong force promoting 3PP in the real world. Of course, however, the absolute amount of HOP observed in a given situation should depend enormously on many contextual factors (e.g., the nature and severity of the original transgression, the cost and potential impact of punishment, the social context, etc.). Moreover, work investigating repeated punishment with the possibility of communication has revealed that initial punishments are sometimes symbolic, increasing subsequently if behavior does not change (Masclet, Noussair, Tucker, & Villeval, 2003; Ostrom, Walker, & Gardner, 1992), and so punishment in the one shot interactions here may underestimate the use of HOP in the real world. Thus, our experiments cannot (and were not designed to) support any general inferences regarding the absolute strength of HOP in daily life. Rather, our experiments reveal something about the psychology that subjects bring from daily life into our experiments: it is more likely to encourage HOP in the context of 3PP (where HOP is a relatively stronger force promoting social welfare) than in the context of 2PP.

Also, our functional argument does not imply that higher-order punishers are *explicitly* motivated by the goal of promoting social welfare. In general, functional

Electronic copy available at: https://ssrn.com/abstract=3080990

explanations for behavior neither require nor specifically predict equivalent proximate motives (Frank, 1988; Martin & Cushman, 2016a). And indeed, in the context of punishment, much research suggests that it is unlikely that individuals punish with such consequentialist considerations in mind (Carlsmith, 2008; Carlsmith & Darley, 2008; Carlsmith, Darley, & Robinson, 2002; Darley, Carlsmith, & Robinson, 2000; Eriksson, Andersson, & Strimling, 2016, 2017; Eriksson, Strimling, et al., 2017). Instead, as discussed above and consistent with our proposal that 3PP is perceived as normative, a proximate mechanism for HOP may be the view that impartial third parties have a role or duty to punish (Eriksson et al., 2013; Strimling & Eriksson, 2014). Our results in Experiment 4 provide evidence in favor of this idea, demonstrating that HOP of 3PP is greatest when the third party plays a particular role (that of an authority or a friend of the victim), though we nevertheless find HOP of 3PP even when the third party is a stranger to the victim.

5.3 Outstanding questions

Our findings highlight the need for work to further characterize the contexts in which people see punishment as normative. Evidence suggests that perceived obligations to punish are likely to vary considerably across contexts (Eriksson, Andersson, et al., 2017; Eriksson, Strimling, et al., 2017), and it is not generally the case that when any victim is harmed, all third-party observers rush to personally carry out punishment of the harm-doer. In the experiments that we designed, third-party observers had relatively unique knowledge and opportunity to punish a harm-doer, and the victim was not present and able to punish for themselves. These factors may increase the extent to which 3PP is perceived as normative or obligatory.

Finally, an open and interesting question is the degree to which there are individual differences in the tendency to view 3PP as injunctively normative and engage in HOP. While our analyses about the frequency of HOP begin to shed light on how often such behavior occurs, future work should investigate which traits correlate with viewing 3PP as required and being willing to sanction those who fail to punish on behalf of others.

5.4 Conclusion

We find HOP directed at third-party observers who fail to punish. Furthermore, we consistently observe that greater sanctions are imposed upon those failing to engage in 3PP than those failing to engage in 2PP. These results suggest that, at least in the contexts investigated here, some people see 3PP as injunctively normative, and that 3PP is seen as more injunctively normative than 2PP. They also suggest that when 2PP is observed, it is likely not motivated by social pressure to punish, but instead by more direct benefits, such as its value as a deterrent. Higher-order punishment may be more “prosocial” in the context of 3PP, serving to motivate unaffected observers to punish selfishness, and thus helping to promote both the punishment of non-cooperators and, ultimately, cooperation itself. Our study thus sheds light on the psychological mechanisms underlying punishment, its social normative basis, its enforcement by higher-order punishment, and the potential role of punishment in promoting human social welfare.

Electronic copy available at: https://ssrn.com/abstract=3080990

Acknowledgements

We thank Max Krasnow, Ryan Miller, Jonathan Phillips, Felix Warneken, the SHAME writing group and members of the Moral Psychology Research Laboratory for their advice and assistance. FC was supported by grant N00014-19-1-2025 from the Office of Naval Research and grant 61061 from the John Templeton Foundation.

Author contributions

JWM, JJJ, DGR and FC designed research; JWM and JJJ performed research; JWM and JJJ analyzed data; JWM, JJJ, DGR and FC wrote the paper.

Competing financial interests

The authors declare no competing financial interests.

Electronic copy available at: https://ssrn.com/abstract=3080990

References

Balafoutas, L., & Nikiforakis, N. (2012). Norm enforcement in the city: A natural field experiment. *European Economic Review*, *56*(8), 1773–1785. https://doi.org/10.1016/j.euroecorev.2012.09.008

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences of the United States of America*, 1–4. https://doi.org/10.1073/pnas.1413170111

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological Bulletin*, *137*(4), 594–615. https://doi.org/10.1037/a0023489

Balliet, D., & Van Lange, P. A. M. (2013). Trust, Punishment, and Cooperation Across 18 Societies: A Meta-Analysis. *Perspectives on Psychological Science*, *8*(4), 363–379. https://doi.org/10.1177/1745691613488533

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344. https://doi.org/10.1016/j.evolhumbehav.2006.01.003

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.

Bone, J. E., & Raihani, N. J. (2015). Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2015.02.002

Brañas-Garza, P., Espín, A. M., Exadaktylos, F., & Herrmann, B. (2014). Fair and unfair punishers coexist in the Ultimatum Game. *Scientific Reports*, *4*, 1–4. https://doi.org/10.1038/srep06025

Electronic copy available at: https://ssrn.com/abstract=3080990

Brandt, H., Hauert, C., & Sigmund, K. (2006). Punishing and abstaining for public goods. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 495–497. https://doi.org/10.1073/pnas.0507229103

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*(5), 930–940. https://doi.org/10.1016/j.neuron.2008.10.016

Carlsmith, K. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, *21*(2), 119–137. https://doi.org/10.1007/s11211-008-0068-x

Carlsmith, K., & Darley, J. (2008). Psychological aspects of retributive justice. *Advances in Experimental Social Psychology*, *40*(7), 193–236. https://doi.org/10.1016/S0065-2601(07)00004-4

Carlsmith, K., Darley, J., & Robinson, P. (2002). Why do we punish?: Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284–299. https://doi.org/10.1037/0022-3514.83.2.284

Carpenter, J., & Matthews, P. H. (2009). What norms trigger punishment? *Experimental Economics*, *12*(3), 272–288. https://doi.org/10.1007/s10683-009-9214-z

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, *9*(3), 265–279. https://doi.org/10.1007/s10683-006-9127-z

Cinyabuguma, M., Page, T., & Putterman, L. G. (2004). On Perverse and Second-Order Punishment in Public Goods Experiments with Decentralized Sanctioning. *SSRN Electronic Journal*, (October 2016). https://doi.org/10.2139/ssrn.724228

Electronic copy available at: https://ssrn.com/abstract=3080990

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

Cushman, F. (2015). Punishment in Humans: From Intuitions to Institutions. *Philosophy Compass*, *10*(2), 117–133. https://doi.org/10.1111/phc3.12192

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a “trembling hand” game. *PloS One*, *4*(8), e6699. https://doi.org/10.1371/journal.pone.0006699

Darley, J., Carlsmith, K., & Robinson, P. (2000). Incapacitation and just deserts as motives for punishment. *Law and Human Behavior*, *24*(6), 659–683.

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254–1258. https://doi.org/10.1126/science.1100735

Eriksson, K., Andersson, P. A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations*, *19*(2), 152–168. https://doi.org/10.1177/1368430215583519

Eriksson, K., Andersson, P. A., & Strimling, P. (2017). When is it appropriate to reprimand a norm violation? The roles of anger, behavioral consequences, violation severity, and social distance. *Judgment and Decision Making*, *12*(4), 396–407.

Eriksson, K., Strimling, P., Andersson, P. A., Aveyard, M., Brauer, M., Gritskov, V., … Yamagishi, T. (2017). Cultural Universals and Cultural Differences in Meta-Norms about Peer Punishment. *Management and Organization Review*, (September), 1–20. https://doi.org/10.1017/mor.2017.42

Eriksson, K., Strimling, P., & Ehn, M. (2013). Ubiquity and efficiency of restrictions on

Electronic copy available at: https://ssrn.com/abstract=3080990

informal punishment rights. *Journal of Evolutionary Psychology*, *11*(1), 17–34. https://doi.org/10.1556/JEP.11.2013.1.3

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4). https://doi.org/10.1016/j.tics.2004.02.007

Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/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), 1–25. https://doi.org/10.1007/s12110-002-1012-7

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a

FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J. J., & Phelps, E. a. (2014). Fairness violations elicit greater punishment on behalf of another than for oneself. *Nature Communications*, *5*, 5306. https://doi.org/10.1038/ncomms6306

Fowler, J. (2005). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences*, *102*(19), 7047–7049. https://doi.org/10.1073/pnas.0500938102

Frank, R. (1988). *Passions within reason: The strategic role of the emotions.* New York: Norton.

Fu, T., Ji, Y., Kamei, K., & Putterman, L. (2017). Punishment can support cooperation even when punishable. *Economics Letters*, *154*, 84–87. https://doi.org/10.1016/j.econlet.2017.01.016

Electronic copy available at: https://ssrn.com/abstract=3080990

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*(5907), 1510. https://doi.org/10.1126/science.1164744

Gaechter, S. (2014). Human Pro-Social Motivation and the Maintenance of Social Order.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, *24*, 153–172. https://doi.org/10.1016/S1090-5138(02)00157-5

Gordon, D. S., Madden, J. R., & Lea, S. E. G. (2014). Both loved and feared: third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PloS One*, *9*(10), e110045. https://doi.org/10.1371/journal.pone.0110045

Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition*, *133*(1), 97–103. https://doi.org/10.1016/j.cognition.2014.06.001

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed Models —The R package pbkrtest. Journal of Statistical Software. https://doi.org/10.18637/jss.v059.i09

Henrich, J. (2006). Cooperation, Punishment, and the Evolution of Human Institutions. *Science*, *312*(5770), 60–61. https://doi.org/10.1126/science.1126398

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 (New York, N.Y.)*, *327*(5972), 1480–1484. https://doi.org/10.1126/science.1182238

Electronic copy available at: https://ssrn.com/abstract=3080990

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Ziker, J. (2006). Costly punishment across human societies. *Science (New York, N.Y.)*, *312*(5781), 1767–1770. https://doi.org/10.1126/science.1127333

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367. https://doi.org/10.1126/science.1153808

Hilbe, C., Traulsen, A., Rohl, T., & Milinski, M. (2014). Democratic decisions establish stable authorities that overcome the paradox of second-order punishment. *Proceedings of the National Academy of Sciences*, *111*(2), 752–756. https://doi.org/10.1073/pnas.1315273111/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1315273111

Horita, Y. (2010). Punishers May Be Chosen as Providers But Not as Recipients. *Letters on Evolutionary Behavioral Science*, *1*(1), 6–9. https://doi.org/10.5178/lebs.2010.2

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. https://doi.org/10.1038/nature16981

Jordan, J. J., & Rand, D. G. (2017). Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology*, *421*, 189–202. https://doi.org/10.1016/j.jtbi.2017.04.004

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology*, *95*(4), 826–842. https://doi.org/10.1037/a0011381

Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What are

Electronic copy available at: https://ssrn.com/abstract=3080990

punishment and reputation for? *PloS One*, *7*(9), e45662. https://doi.org/10.1371/journal.pone.0045662

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking Under the Hood of Third-Party Punishment Reveals Design for Personal Benefit. *Psychological Science*, *27*(3), 405–418. https://doi.org/10.1177/0956797615624469

Kriss, P. H., Weber, R. A., & Xiao, E. (2015). Turning a Blind Eye, But Not the Other Cheek: On the Robustness of Costly Punishment, (January).

Kurzban, R., Descioli, P., & Obrien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models [Computer software manual].

Leibbrandt, A., & López-Pérez, R. (2012). An exploration of third and second party punishment in ten simple games. *Journal of Economic Behavior & Organization*, *84*(3), 753–766. https://doi.org/10.1016/j.jebo.2012.09.018

Martin, J. W., & Cushman, F. (2015). To Punish or to Leave: Distinct Cognitive Processes Underlie Partner Control and Partner Choice Behaviors. *PLOS ONE*, *10*(4), e0125193. https://doi.org/10.1371/journal.pone.0125193

Martin, J. W., & Cushman, F. (2016a). The adaptive logic of moral luck. In J. Sytsma & W. Buckwalter (Eds.), *The Blackwell Companion to Experimental Philosophy* (pp. 190–202). Malden, MA: Wiley-Blackwell.

Martin, J. W., & Cushman, F. (2016b). Why we forgive what can’t be controlled. *Cognition*, *147*, 133–143. https://doi.org/10.1016/j.cognition.2015.11.008

Electronic copy available at: https://ssrn.com/abstract=3080990

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. *American Economic Review*, *93*(1), 366–380. https://doi.org/10.1257/000282803321455359

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences*, *108*(28), 11375–11380. https://doi.org/10.1073/pnas.1105604108

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *The Behavioral and Brain Sciences*, *36*(1), 1–15. https://doi.org/10.1017/S0140525X11002160

Mendoza, S. a., Lane, S. P., & Amodio, D. M. (2014). For Members Only: Ingroup Punishment of Fairness Norm Violations in the Ultimatum Game. *Social Psychological and Personality Science*, *5*(6), 662–670. https://doi.org/10.1177/1948550614527115

Morris, A., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Evolution of flexibility and rigidity in retaliatory punishment. *Proceedings of the National Academy of Sciences*, *114*(39), 10396–10401. https://doi.org/10.1073/pnas.1704032114

Nelissen, R. M. A. (2008). The price you pay: cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, *29*(4), 242–248. https://doi.org/10.1016/j.evolhumbehav.2008.01.001

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a Sword: Self-Governance Is Possible. *American Political Science Review*, *86*(2), 404–417. https://doi.org/10.2307/1964229

Patil, I., Dhaliwal, N. A., & Cushman, F. (2018). Reputational and cooperative benefits

Electronic copy available at: https://ssrn.com/abstract=3080990

of third-party compensation. *Psyarxiv*, 1–56.

Raihani, N. J., & Bshary, R. (2015a). The reputation of punishers. *Trends in Ecology & Evolution*, 1–6. https://doi.org/10.1016/j.tree.2014.12.003

Raihani, N. J., & Bshary, R. (2015b). Third-party punishers are rewarded - but third-party helpers even more so. *Evolution*, (2013), 2–4. https://doi.org/10.1111/evo.12637.This

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology & Evolution*, *27*(5), 288–295. https://doi.org/10.1016/j.tree.2011.12.004

Strimling, P., & Eriksson, K. (2014). Regulating the regulation: Norms about punishment. *Social Dilemmas: Punishment and Rewards*, 52–69. https://doi.org/10.1016/j.celrep.2011.1011.1001.7.

Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. L., Ginther, M. R., … Marois, R. (2014). Corticolimbic gating of emotion-driven punishment. *Nature Neuroscience*, *17*(August), 1270–1275. https://doi.org/10.1038/nn.3781

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. https://doi.org/10.1086/406755

Whitson, J. a., Wang, C. S., See, Y. H. M., Baker, W. E., & Murnighan, J. K. (2015). How, when, and why recipients and observers reward good deeds and punish bad deeds. *Organizational Behavior and Human Decision Processes*, *1*(October 2017), 84–95. https://doi.org/10.1016/j.obhdp.2015.03.006

Zhou, Y., Jiao, P., & Zhang, Q. (2016). Second-party and third-party punishment in a public goods experiment. *Applied Economics Letters*, 0–29.

Electronic copy available at: https://ssrn.com/abstract=3080990