Organizational Behavior and Human Decision Processes 164 (2021) 27–51

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Reputational and cooperative benefits of third-party compensation

Nathan A. Dhaliwal [a,1,*], Indrajeet Patil [b,1], Fiery Cushman [c,2]

[a] UBC Sauder School of Business, University of British Columbia, Vancouver, Canada
[b] Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany
[c] Department of Psychology, Harvard University, Cambridge, United States

ARTICLE INFO

Keywords:
Morality
Punishment
Justice
Signalling
Prosociality

ABSTRACT

Although third-party punishment helps sustain group cooperation, might victim compensation provide third parties with superior reputational benefits? Across 24 studies (N = 21,296), we provide a comprehensive examination of the consequences of the choice between punishment and compensation. What do people infer from, and how do they respond to, the choice of punishment versus compensation? Across various contexts ranging from economic games, to workplace injustice, to people's own personal experience of witnessing third-party responses in their daily lives, we find that compensating victims leads to greater reputational and partner choice benefits relative to punishing perpetrators. In fact, even people who themselves prefer to punish still prefer social partners who compensate. We also find that the signal that is sent via third-party compensating is an honest signal—people who choose to compensate rather than punish score lower on measures of trait Machiavellianism, narcissism, and psychopathy. Furthermore, we find that the personal decision of whether to compensate or punish is influenced by both injunctive and descriptive norms. These findings provide an extensive analysis of the causes and consequences of third-party responding to moral violations.

1. Introduction

Third-party punishment is a focal point of research across the behavioral sciences (Boyd & Richerson, 1992; Boyd, Gintis, Bowles, & Richerson, 2003; Boyd et al., 2010; Henrich & Boyd, 2001). In its canonical form, third-party punishment occurs when a *perpetrator* harms a *victim*, and therefore some *third party* punishes the perpetrator. This behavior is interesting for two reasons. First, third-party punishment offers a potential solution to a fundamental mystery about humans: Why do we, alone, cooperate within large-scale groups of genetically unrelated individuals (Boyd & Richerson, 1992, 2005; Henrich, 2015; Henrich & Henrich, 2007; Richerson & Boyd, 2005; Richerson & Boyd, 1998)? Theoretical and empirical studies show that this kind of cooperative behavior can be explained in the presence of third-party punishment (Balliet et al., 2011; Fehr & Gächter, 2000; Mathew & Boyd, 2011). Second, however, third-party punishment itself poses a conundrum: because its most obvious benefit is shared by a group, but the costs are born by individuals, it is not immediately clear how it could be favored by natural selection operating at the individual level.

One solution to this puzzle is that, by intervening, a third party could gain an improved reputation, specifically that punishing can signal trustworthiness (Barclay, 2006; Jordan, Hoffman, Bloom, & Rand, 2016; Nelissen, 2008; Raihani & Bshary, 2015a). Importantly, despite some well-known studies demonstrating one-shot anonymous third-party punishment, these accounts suppose that third-party punishment is adapted to contexts characterized by repeated and non-anonymous interactions. This potential solution to the puzzle of third-party punishment—moral reputation—is our point of departure. Most prior research supporting this model has focused on the reputational benefits of third-party punishment versus *no third-party response at all*. In principle, however, there are many ways a third party could intervene following a transgression beyond merely punishing or doing nothing, like the possibility of compensation, in which third parties bestow a benefit upon the victim of a norm violation.

Against this background, recent studies have shown that third parties often do prefer to compensate victims rather than punish transgressors when given a forced choice (FeldmanHall, Sokol-Hessner, Van Bavel, & Phelps, 2014; van Doorn, Zeelenberg, & Breugelmans, 2018). Moreover,

Abbreviations: DG, Dictator Game; TG, Trust Game.

* Corresponding author.
E-mail address: nathan.dhaliwal@sauder.ubc.ca (N.A. Dhaliwal).

[1] These authors contributed equally to this work and share first authorship.

[2] FC was supported by grant 61061 by the John Templeton Foundation.

https://doi.org/10.1016/j.obhdp.2021.01.003
Received 13 September 2019; Received in revised form 14 December 2020; Accepted 6 January 2021

this appears to be an adaptive choice, because several studies now also show that choosing to compensate, rather than to punish, improves one's moral reputation. Specifically, it has been demonstrated that compensating signals more trustworthiness than punishing (Jordan, Hoffman, Bloom, & Rand, 2016), that compensators are chosen more often as cooperation partners (Heffner & FeldmanHall, 2019), that people are more approving of third-party compensators and expect third parties to compensate rather than punish (Li et al., 2018), and that compensators are monetarily rewarded more than punishers (Raihani & Bshary, 2015b).

This finding is notable because, while third-party punishing and third-party compensating are both laudable third-party responses to injustice, only third-party punishing serves to deter wrongdoing and stabilize cooperation within a group (Balliet et al., 2011; Fehr & Gächter, 2000; Mathew & Boyd, 2011). Thus, it might seem likely that punishment would garner greater reputational benefits than compensation. Why, instead, does compensation provide superior reputational benefits?

Our answer is that compensation is a costly, honest signal of prosocial tendencies, and a more reliable signal than punishment is. To build this case we carry out a comprehensive analysis of (1) the reputational benefits afforded to third-party responses, (2) whether they are correctly anticipated by third-party actors, and (3) whether they are grounded in reality. Having established this empirical case, in the general discussion we take up the question of why compensation is more reliable, pursuing this from the perspective of costly signaling theory.

Our approach is organized around four main questions. First, what are the specific reputational consequences of compensation versus punishment? We probe a much wider variety of reputational consequences than has been previously explored. Specifically, we examine perceived morality, trustworthiness, generosity, warmth, competence, emotionality, logicality, praiseworthiness, blameworthiness, psychopathy, integrity, benevolence, desire to work for and be loyal to, and finally expectations for what a good person would do. We also ask whether engaging in compensation and punishment together can produce superior reputational benefits to engaging in either alone. We find that they may not, indicating that they may "compete" for a common reputational influence. We replicate and extend prior studies by showing that compensating leads one to be chosen more often as a cooperation partner and whether compensators receive more money in an economic exchange. We also examine whether people infer that punishers may be riskier cooperation partners by testing whether punishers are seen as harbouring greater psychopathic tendencies relative to compensators.

Second, do people accurately anticipate the reputational consequences of compensating versus punishing? We find that they do. Moreover, they shift their decision to compensate versus punish depending on which kind of response is most prevalent in their social group and which response will provide superior reputational benefits, and accurately predict that a victim would prefer compensation to punishment on their behalf. Collectively, these findings establish that people have the knowledge required to make choices about compensation versus punishment based on their reputational consequences.

Third, is the perception that compensators are more prosocial than punishers grounded in reality? We first examine whether those who choose compensation behave more cooperatively in economic games. Next, we examine how the same groups differ in their Dark Triad of personality traits (Paulhus & Williams, 2002), namely, Machiavellianism (Christie & Geis, 1970), narcissism, and psychopathy (Jones & Paulhus, 2014).

Fourth, how generalizable are these effects across a wide variety of contexts, including people's actual lives? This is important as previous field research has shown that victim preferences for compensation over punishment may differ depending on the type of violation that the victim endured; specifically, compensation may be seen as more desirable when addressing minor violations rather than more major moral violations which trigger greater moral outrage (Darley & Pittman, 2003; Reb et al., 2006). Furthermore, Darley and Pittman (2003) propose that intentional harms are more likely to trigger high moral outrage leading to third-party punishment whereas negligent or accidental harms are likely to trigger less moral outrage leading to third-party compensation. Thus, we examine whether observer perceptions of compensators over punishers differs across various types of violations. For example, we test if third-party compensators are preferred when responding to norm violations in economic games, workplace and societal vignettes, and when people recall witnessing norm violations in their daily lives. We also test whether punishers are seen as better suited for certain leadership roles and whether people shift their preference from compensating the victim to punishing the perpetrator depending on injunctive norms.

As a further test of generalizability, across all our studies we also investigate an important potential source of individual differences: Does a person's own choice to punish (or compensate) predict how they update the reputation of somebody else who also chooses to punish (or compensate)? This question is not addressed in previous research. By asking it, we can test whether there is a "generally preferred" form of third-party intervention or whether, instead, people simply prefer for others to respond to third parties in whatever manner they choose for themselves. This is important because if compensators are preferred even amongst those who prefer to punish, such would suggest a particularly strong selection pressure favoring third-party compensation.

Across the majority of the studies reported in this paper we ask participants to make judgments or indicate their preferred behavior in response to hypothetical economic games. This allowed us to explore a multitude of research questions about third-party justice behavior while avoiding the inherent idiosyncratic nature of contextualized vignette studies, thus aiding in the potential generalizability of these the findings. Given that we sought to test a vast range of research questions, we were forced to choose a specific context within to test most of these questions. We chose an economic game given that it is free of rich contextual details and thus may stand the best chance providing generalizable results. But, of course, using this method comes with certain costs, particularly concerns with regards to external validity as well concerns about the hypothetical and non-incentivised nature of these experiments (Patil, et al., 2014; Patil, et al., 2017; Teper, Tullett, Page-Gould, & Inzlicht, 2015).

To address some of these concerns, we examine in Part 6 of this paper whether our primary findings replicate across multiple contextualized vignettes as well as when participants are asked to reflect on their actual lived experiences of witnessing third-party responders. Specifically, we examine whether compensators gain greater reputational benefits than punishers when they are addressing: a Ponzi scheme, property theft, genocide, domestic violence, verbal harassment, workplace injustice, and hundreds of different experiences from participants' own personal lives.

2. Part 1: Reputational benefits

We begin by examining the reputational outcomes of third-party punishing and compensating in an economic game. Adapting several approaches from prior research, we measure perceived trustworthiness, ethicality, and generosity (Eisenbruch & Roney, 2017; Everett, Pizarro, & Crockett, 2016; Jordan, Hoffman, Bloom, & Rand, 2016; Smith & Bird, 2000) inferences about a third party's warmth and competence (Judd, James-Hawkins, Yzerbyt, & Kashima, 2005) and impressions about whether a third party's decision was governed more by affective or cognitive processing (Epstein et al., 1996). Each of these judgments contribute to the overall impression of a person's character (Everett,

Pizarro, & Crockett, 2016; Rom et al., 2017).

Along with judgments about the morality of actors (i.e., their character), we also examine the perceived praiseworthiness and blameworthiness[3] of the action of punishing and compensating. This reflects an important conceptual and empirical distinction between *act-based* versus *person-centred* approach to moral judgment (Cushman, 2015a; Landy and Uhlmann, 2018; Robinson et al., 2017; Uhlmann, Pizarro, & Diermeier, 2015).

3. General methods for reputational benefits studies (1a–e)

3.1. Ethics statement

All studies were carried out on Amazon Mechanical Turk and were approved by the Ethics Committee of Harvard University under the umbrella protocol (IRB14-2016). We conducted all studies and collected all data described in this paper. None of the data comes from previously developed datasets from third parties.

3.2. Procedure

Across the majority of the studies in the paper we employed variations of a hypothetical third-party punishing and compensating game. Participants began each experiment by being presented with the instructions for this game. The basic instructions for the game were as follows:

> "In this interaction there are three people: Person A, Person B, and Person C.
> Person A and Person B each start with 100 cents.
> First Person A was asked to make a choice
> 1) Person A could choose to take or not take from Person B.
> 2) If A takes, Person B loses 50 cents and A gains 50 cents.
> 3) Afterwards, Person C is given 10 cents and can spend that money to cause A to lose money or to cause B to gain money. Whatever money Person C does not spend, Person C gets to keep for him/herself.
> For every 1 cent Person C spends, Person C can cause:
> Person A to lose 5 cents
> Or
> Person B to gain 5 cents
> C must choose whether to cause A to lose cents or cause B to gain cents; C cannot do both. Person B is passive in this interaction and makes no decisions."

Along with the instructions, participants were also shown a figure depicting the interaction (see Fig. 1a).

Unless and otherwise stated, participants were also asked how they would respond as Person C, which signified their own *personal preference* in response to an observed norm violation. Across all studies, personal preference was measured using a forced dichotomous choice of whether participants personally preferred the option to punish or compensate. When there were multiple questions, the order of the questions was randomized across participants.

3.3. Exclusion criteria

After being presented with the instructions for the game, participants were asked (3 or 4, depending on the study) comprehension questions to ensure they understood the rules. If a participant failed to answer any one of the comprehension questions correctly, they were not allowed to proceed with the experiment. This stringent criterion is commonly used in economic games to ensure that all participants who take part in the experiment fully understand how the game works (Capraro et al., 2018). Additionally, participants who took unusually short or long amounts of time to complete the task (z scores for duration ≥ 3) were also removed on the assumption that they had not devoted full and undivided attention to the task.

3.4. Data analysis

The majority of analyses was carried out in the R programming language. Data was modelled using General Linear Model. For the sake of brevity, demographic details for all studies (age summary statistics and gender breakdown) are provided in Table 1.

3.5. Data visualization

For all plots relevant for group or condition comparisons, we show a mix of violin plot (which displays the shape of the variable distribution) and box-and-whisker plot (where the box is split in the middle by median and bounded by first and third quartiles of the distribution, while the whiskers show minimum and maximum values excluding outliers) along with the raw data points jittered to avoid overlap. In addition, the red dot in all plots signifies mean values. For brevity, many statistical parameters are included in the figures rather than the main text (an approach adopted in the R package *ggstatsplot* (Patil, 2018)).

3.6. Data availability

All data and scripts are available at Open Science Framework: https://osf.io/yhbrc/.

4. Study 1a

To begin, we examined the inferences people make about the trustworthiness and ethicality of third-party punishers versus compensators.

4.1. Methods

A total of 1002 participants were randomly assigned to read that, in response to A taking from B, C had chosen to either 1) punish A by paying to remove 50 cents from A (henceforth referred to as the "punisher" condition), or 2) compensate B by paying to give 50 cents to B (henceforth referred to as the "compensator" condition). After reading whether C chose to punish or compensate, participants responded using a 7-point scale regarding their impressions of how trustworthy and how moral Person C is. Since these two ratings were highly correlated ($r = 0.9$), we averaged them to form a single moral character index (hereafter, "character"). Participants were also asked how they would respond as Person C, which signified their own *personal preference* in response to the norm violation.

4.2. Results

The results revealed a main effect of condition, preference, and an interaction between condition and preference (for full statistical details, see Table 2). As can be seen from Fig. 2a, irrespective of whether they

[3] We would like to note that we are not arguing that the act of compensating or punishing as a third-party in and of itself is blameworthy (or praiseworthy), but rather by choosing a certain means of intervention, one is forgoing the benefits that are brought about by the alternative intervention strategy. For example, by choosing to punish, one is choosing to not provide the victim with any material benefits. On the other hand, by choosing to compensate, one is letting the perpetrator get away with the wrongdoing. It may be the case that people's judgments of the character of compensators aligns with their judgments of the act of compensating, such that compensating is seen as less blameworthy and worthier of praise. On the other hand, it may be the case that compensators who also allow the perpetrator to get away with the wrongdoing are seen as less praiseworthy and worthier of blame.

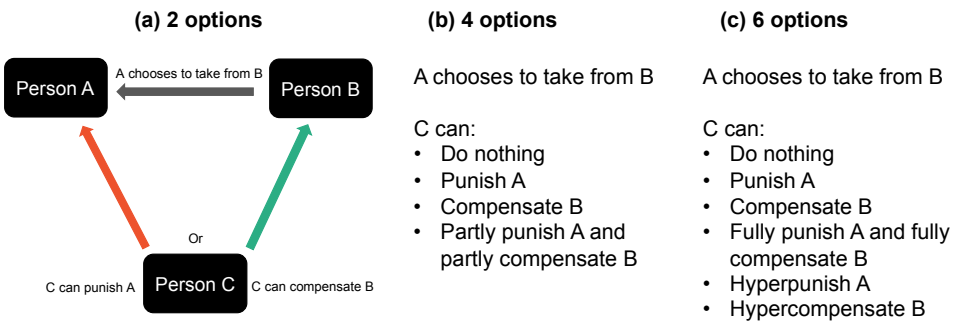

Fig. 1. (*a*) Schematic illustration of the Third-party punishing and compensating game with two options (Studies 1a–f, 3a–f, 5a, 5b). (*b*) Variant of this game that includes four options (Studies 2a, 2b, 3g). (*c*) Variant of this game that includes six options (Study 2c).

personally preferred compensation or punishment, they rated compensators to have a more positive moral character than punishers, although the effect size was significantly larger for people who personally preferred compensation. It is also interesting to note that irrespective of whether participants preferred to compensate or punish, they rated third-party punishers above the midpoint on the scale, indicating a positive impression of third-party punishers. Thus, it is not the case that third-party punishers are viewed negatively, rather that people show a particularly strong preference for third-party compensators.

5. Study 1b

Study 1b attempted a direct replication of Study 1a. A total of 981 participants participated in this study. The only difference was that the presentation order of dependent measures (trustworthiness, morality, and personal preference) was randomized. Replicating Study 1a, compensators were judged to have a superior moral character than punishers and this effect was stronger in people who personally preferred compensation (see Fig. 2b; for full statistical details, Table 2).

6. Study 1c

While Study 1a and 1b focused on the perceived ethicality and trustworthiness of the third party, Study 1c focuses on the perceived generosity of the third party. In the third-party punishing and compensating game, the third party is being equally generous by punishing or compensating (in each case they are spending 10 cents). But, if victim compensation signals better moral character, it is possible that people also perceive compensators to be more generous as generosity is closely related to virtuous behavior.

6.1. Methods

A total of 989 participants took part in this study. The procedure was the same as in Study 1a-b except that after reading whether C chose to punish or compensate, participants responded using a 7-point scale regarding their impressions of how generous Person C is.

6.2. Results

The results revealed a main effect of condition, preference, and an interaction between condition and preference (for full statistical details, see Table 2). As can be seen from Fig. 2c, irrespective of whether they personally preferred compensation or punishment, they rated compensators to be more generous than punishers, although the effect size was significantly larger for people who personally preferred compensation.

7. Study 1d

Humans infer traits going far beyond morality, trustworthiness, and generosity. Therefore, in the next study, we ask whether observing third party intervention supports selective inferences of traits relevant to morality and cooperation (trustworthiness, morality, generosity, warmth, etc.), or whether it instead supports more general trait inferences (e.g., competence, logical reasoning, emotionality, etc.) (Epstein et al., 1996; Judd, James-Hawkins, Yzerbyt, & Kashima, 2005). Moreover, research has shown that although morality and warmth may seem interchangeable, they are in fact distinct constructs (Brambilla & Leach, 2014; Goodwin, Piazza, & Rozin, 2014). By measuring both perceived morality and warmth, we could attempt to gain a more fine-grained understanding about what third-party compensation signals.

7.1. Methods

A total of 805 participants were randomly assigned to either punisher or compensator condition. After reading whether C chose to punish or compensate, participants were asked how much they thought C's decision was based on feelings and emotions and how much they thought C's decision was based on logical reasoning. Next, participants were asked, using a 7-point scale, to indicate their impressions of Person C's:

(i) warmth (measured using the items: warm, good-natured, tolerant, sincere; Cronbach's $\alpha = 0.90$)
(ii) competence (measured using the items: competent, confident, independent, competitive, intelligent; Cronbach's $\alpha = 0.79$),
(iii) trustworthiness,
(iv) morality.

Based on the correlation matrix, morality, trustworthiness and warmth were collapsed into a single measure of character (Cronbach's $\alpha = 0.87$; see Supplementary Fig. 1).

7.2. Results

Replicating results from Study 1a-b, compensators were judged to have a better moral character, and this effect was much stronger for people who personally preferred compensation than those who preferred punishment (Fig. 3b). Crucially, however, the same directional effect was observed in both groups, indicating a general preference for compensators.

A different pattern emerged with our additional, non-moral measures of trait inferences. Here, instead, we observe a set of crossover interactions in which the directional effects of trait ascription to punishment versus compensation depends on whether participants themselves would choose to punish or compensate. First, participants with a personal preference for punishment found punishers to be more competent, while participants with a personal preference for compensation trended in the opposite direction. Thus, it appears that while compensating signals greater moral character, punishment may signal higher levels of competence. Second, participants with a personal

Table 1
Summary for age and breakdown by gender for all studies. *Abbreviations*: n-sample size, sd- standard deviation.

study	gender	mean	sd	n
1a	male	34.08	10.43	414
1a	female	36.18	12.11	586
1a	other	25	2.83	2
1b	male	31.77	9.35	406
1b	female	35.45	11.45	569
1b	other	30.5	14.2	6
1c	male	36.21	12.41	418
1c	female	36.96	12.33	566
1c	other	27.2	4.09	5
1d	male	34.13	10.73	373
1d	female	35.75	11.61	428
1d	other	26.2	6.83	5
1e	male	34.4	11.28	656
1e	female	36.41	11.37	971
1e	other	34.67	14.02	6
1f	male	34.81	11.16	93
1f	female	37.53	12.13	109
1f	other	26	NA	1
2a	male	32.47	10.2	447
2a	female	35.75	11.71	558
2a	other	34	10.16	6
2b	male	35.78	10.94	400
2b	female	37.17	11.33	613
2b	other	26	8.49	2
2c	male	33.97	10.41	1127
2c	female	36.37	11.71	1311
2c	other	25.73	2.57	11
3a	male	33.93	9.86	206
3a	female	35.3	11.63	197
3a	other	32	9.9	2
3b	male	34.15	10.27	192
3b	female	36.39	10.86	212
3b	other	25	NA	1
3c	male	33.36	9.34	209
3c	female	34.68	10.61	196
3c	other	22	NA	1
3d	male	34.84	10.64	215
3d	female	36.38	11.46	289
3e	male	37.45	12.36	425
3e	female	37.95	12.39	507
3e	other	23.33	1.53	3
3f	male	34	11.09	381
3f	female	35.94	11.94	592
3f	other	25.75	4.13	8
3g	male	33.85	11.79	466
3g	female	35.57	11.78	595
3g	other	24.5	6.09	6
4a	male	32.98	10.1	86
4a	female	34.03	10.43	117
4a	other	23	NA	1
4b	male	33.88	10.26	274
4b	female	36.83	11.77	326
4b	other	31.2	8.7	5
4c	male	31.96	9.56	158
4c	female	33.75	11.14	240
4c	other	24	NA	1
5a	male	33.54	10.07	428
5a	female	35.84	11.36	574
5a	other	23.5	6.66	4
5b	male	34.84	10.92	364
5b	female	36.1	11.4	583
5b	other	25.75	2.75	4
6a	male	37.43	11.49	1140
6a	female	39.51	12.51	1030
6a	other	27	NA	5
6b	male	35.23	9.54	417
6b	female	38.6	11.30	224
6b	other	28.67	6.11	3
6c	male	37.66	12.37	203
6c	female	39.95	12.66	192
6c	other	28.00	NA	1

preference for compensation found punishers to be more emotionally driven (Fig. 3a), while participants with a personal preference for compensation trended in the opposite direction. Finally, participants with a personal preference for punishment felt that punishment was more based in logic, while participants with a personal preference for compensation showed the opposite effect (see Fig. 3c) (for full statistical details, see Table 2).

Combined with results from Studies 1a-c, the current study establishes that *population-general* positive trait attribution to compensators (versus punishers) is restricted to traits relevant for cooperation (morality, trustworthiness, and warmth). For other traits not directly defined in terms of morality and cooperation—traits such as logical reasoning, emotionality, and competence—people instead exhibit a general tendency to attribute positive traits to people who make the same choices they do.

8. Study 1e

After examining the inferences people make about the quality of various aspects of third party's personality, we next examined the inferences people make about the praiseworthiness and blameworthiness of a third party's behavior. In other words, we complemented our prior studies of *person*-centered judgments with a new study of *act*-centered judgments.

8.1. Methods

Total of 1634 participants participated in this study. In a 2 × 2 between-subjects design, the participants were randomly assigned to either "punisher" or "compensator" condition and indicated using a 7-point scale either how praiseworthy or blameworthy Person C's behavior was.

8.2. Results

The analysis (see Table 2) revealed that, irrespective of personal preference, people found behavior of compensators to be more praiseworthy. On the other hand, only people with a preference for compensation found punishers to be more blameworthy, but no such difference was found for people who personally preferred punishment (Fig. 4). This provides some further support for a population-general preference for compensation; the asymmetry between praise and blame judgments in this regard, however, remains an important topic for further investigation.

9. Study 1f

Studies 1a-e each employ a similar logic: We manipulate information about punishment versus compensation in the stimulus, and then ask participants to draw an inference about character. In this study we reverse the experimental logic, manipulating statements about personal character in the stimuli and then asking participants to draw inferences about the likelihood of punishment versus compensation.

9.1. Methods

A total of 203 participant participated in this study. In a between-subjects design, participants were randomly assigned to one of two conditions where they were asked to imagine either that 1) C is a very trustworthy and moral person, or 2) C is a very untrustworthy and immoral person. Then participants were asked which of the two options (punish A or compensate B) they thought Person C would choose as a response. No personal preference was assessed in this study.

Table 2
Omnibus ANOVA test results for linear models from individual studies.

study	measure	term	F. value	df1	df2	partial omega-squared	conf. low	conf. high	p. value	significance
1a	character	condition	136.82	1	998	0.12	0.08	0.16	1.04E-29	***
	character	preference	19.22	1	998	0.02	0.00	0.03	1.29E-05	***
	character	condition × preference	29.40	1	998	0.03	0.01	0.05	7.40E-08	***
1b	character	condition	146.62	1	977	0.13	0.09	0.17	1.52E-31	***
	character	preference	15.19	1	977	0.01	0.00	0.03	1.04E-04	***
	character	condition × preference	13.27	1	977	0.01	0.00	0.03	2.84E-04	***
1c	generous	condition	330.22	1	985	0.25	0.20	0.30	7.31E-64	***
	generous	preference	0.87	1	985	0.00	0.00	0.00	3.50E-01	ns
	generous	condition × preference	24.44	1	985	0.02	0.00	0.05	9.01E-07	***
1d	affect	condition	2.25	1	802	0.00	−0.01	0.01	0.134212613	ns
	affect	preference	1.29	1	802	0.00	0.00	0.01	0.256901134	ns
	affect	condition × preference	4.88	1	802	0.00	−0.01	0.02	0.027431123	*
	character	condition	228.56	1	802	0.22	0.17	0.26	1.28E-45	***
	character	preference	9.78	1	802	0.01	0.00	0.03	0.001831253	**
	character	condition × preference	14.80	1	802	0.02	0.00	0.04	1.29E-04	***
	cognition	condition	0.38	1	802	0.00	0.00	0.01	0.539199078	ns
	cognition	preference	0.54	1	802	0.00	0.00	0.01	0.462171025	ns
	cognition	condition × preference	21.34	1	802	0.02	0.00	0.05	4.48E-06	***
	competence	condition	1.69	1	802	0.00	−0.01	0.01	0.194070184	ns
	competence	preference	0.08	1	802	0.00	0.00	0.00	0.781066844	ns
	competence	condition × preference	19.12	1	802	0.02	0.00	0.04	1.39E-05	***
1e	blame	condition	19.51	1	810	0.02	0.00	0.04	1.14E-05	***
	blame	preference	6.68	1	810	0.01	0.00	0.02	0.009927778	**
	blame	condition × preference	1.49	1	810	0.00	−0.01	0.01	0.222846107	ns
	praise	condition	77.06	1	815	0.08	0.05	0.12	9.62E-18	***
	praise	preference	10.13	1	815	0.01	0.00	0.03	0.001517176	**
	praise	condition × preference	10.92	1	815	0.01	0.00	0.02	9.92E-04	***
2a	character	condition	71.89	3	995	0.17	0.13	0.23	4.38E-42	***
	character	preference	0.21	3	995	0.00	−0.01	0.01	0.891812656	ns
	character	condition × preference	11.61	9	995	0.09	0.06	0.13	1.80E-17	***
2b	character	condition	53.82	3	999	0.14	0.10	0.17	3.01E-32	***
	character	preference	5.44	3	999	0.01	0.00	0.03	0.00103377	**
	character	condition × preference	5.18	9	999	0.04	0.02	0.07	6.68E-07	***
2c	character	condition	111.94	5	2413	0.18	0.16	0.22	1.29E-106	***
	character	preference	5.15	5	2413	0.01	0.00	0.02	1.06E-04	***
	character	condition × preference	7.74	25	2413	0.06	0.05	0.09	9.83E-27	***
3f	amount	condition	3.74	1	977	0.00	0.00	0.01	0.0534997	ns
	amount	preference	2.71	1	977	0.00	−0.01	0.01	0.09986592	ns
	amount	condition × preference	17.31	1	977	0.02	0.00	0.03	3.46E-05	***
3g	amount	condition	29.22	3	1051	0.07	0.04	0.11	3.78E-18	***
	amount	preference	14.34	3	1051	0.04	0.02	0.06	3.68E-09	***
	amount	condition × preference	3.04	9	1051	0.02	0.00	0.04	0.001329499	**

9.2. Results

Participants expected a trustworthy, moral person to choose to compensate (76%) rather than punish, and overwhelmingly predicted an untrustworthy, immoral person to choose to punish (93%) rather than compensate (see Fig. 5a). It should be noted that the best guess for how an untrustworthy, immoral person would respond to injustice may be that they will choose to do nothing, given that both compensating and punishing are altruistic acts. Nevertheless, when forced to choose between these two altruistic acts, people predict that an immoral person will choose to punish and a moral person will choose to compensate.

10. Part 1: Summary

We began by asking whether people take a more favorable view of third-party punishment, or instead of third-party compensation. We find consistent evidence for the latter: People think that compensators have superior moral character to punishers, they consider compensation a superior moral act to punishment, and they predict that morally virtuous people will compensate more often than they will punish. In most cases these results hold even among participants who would have personally chosen to punish in a third-party role, albeit with lesser magnitude.

11. Part 2: Are reputational effects additive and dependent upon severity?

When responding to a moral violation, a third party isn't limited to the dichotomous choice of either punishing the violator or compensating the victim; a third party could choose to engage both behaviors or could choose simply to not intervene (FeldmanHall et al., 2014). We next examined the reputational repercussions of these alternatives. In addition, if a third party does choose to intervene, they must also decide upon the *magnitude* of the response. Recent research has shown that the decision to intervene as a third party is associated with brain areas responsible for sensitivity to unfairness (e.g. the anterior insula), while the decision regarding the magnitude of the response is associated with brain areas responsible for negative emotion (e.g. the amygdala) (Civai, Huijsmans, & Sanfey, 2019). Might people's reputation vary based not only on the type of response, but also the magnitude of their response? Given that punishing and compensating are both altruistic acts (Hu, Strang, & Weber, 2015; Leliveld, Dijk, & Beest, 2012; O'Gorman et al., 2005), it may seem reasonable to assume that a third party's reputation should improve as the magnitude of the response increases. On the other hand, there may exist an asymmetry such that hypercompensating (paying victims more than the damage they suffered) has positive effects

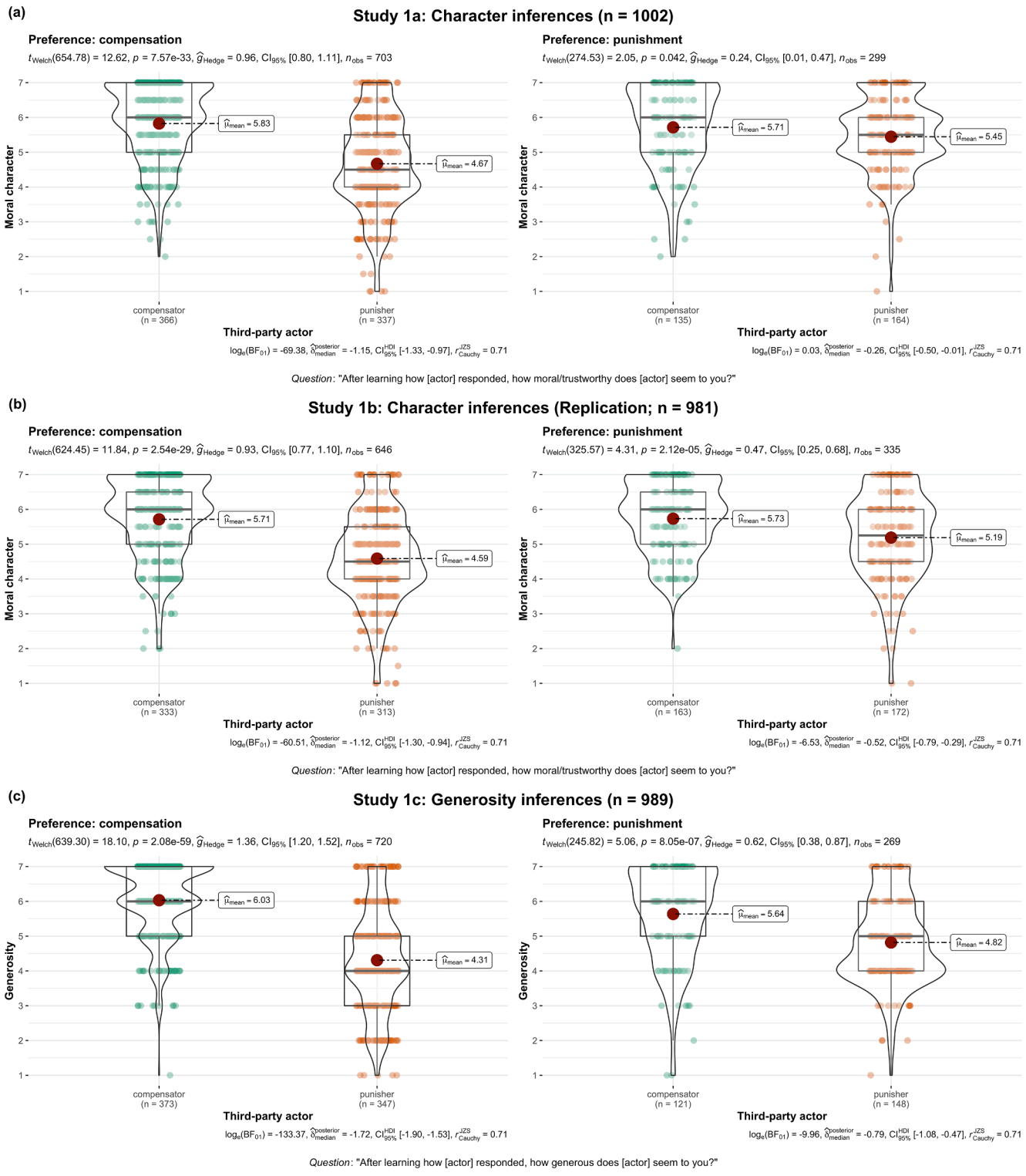

Fig. 2. Inferences about moral character. (*a*) Study 1a showed that people inferred third-party compensators to have a better moral character than punishers, irrespective of their personal preference. (*b*) Results from Study 1a were replicated in Study 1b. (*c*) Inferences about the generosity of a third-party who chose to compensate or punish.

on one's reputation, but hyperpunishing (punishing perpetrators in a manner disproportionate to the damage they caused to the victim) may have adverse effects. This might occur because punishing is an inherently negative sum act (Ohtsuki, Iwasa, & Nowak, 2009) and could be seen by others as being underpinned by negative intentions (Fehr & Rockenbach, 2003; Nakamaru and Iwasa, 2006; Rand et al., 2010). By hyperpunishing, a third party may appear less interested in bringing justice to a situation and more interested in inflicting excessive harm upon the violator. We test whether such an asymmetry exists between hypercompensating and hyperpunishing.

We also examine whether these reputational benefits for compensators are simply a product of the low-stakes nature of the third-party punishing and compensating game (Andersen et al., 2011). It may be the case that when the severity of the unfairness increases, compensators

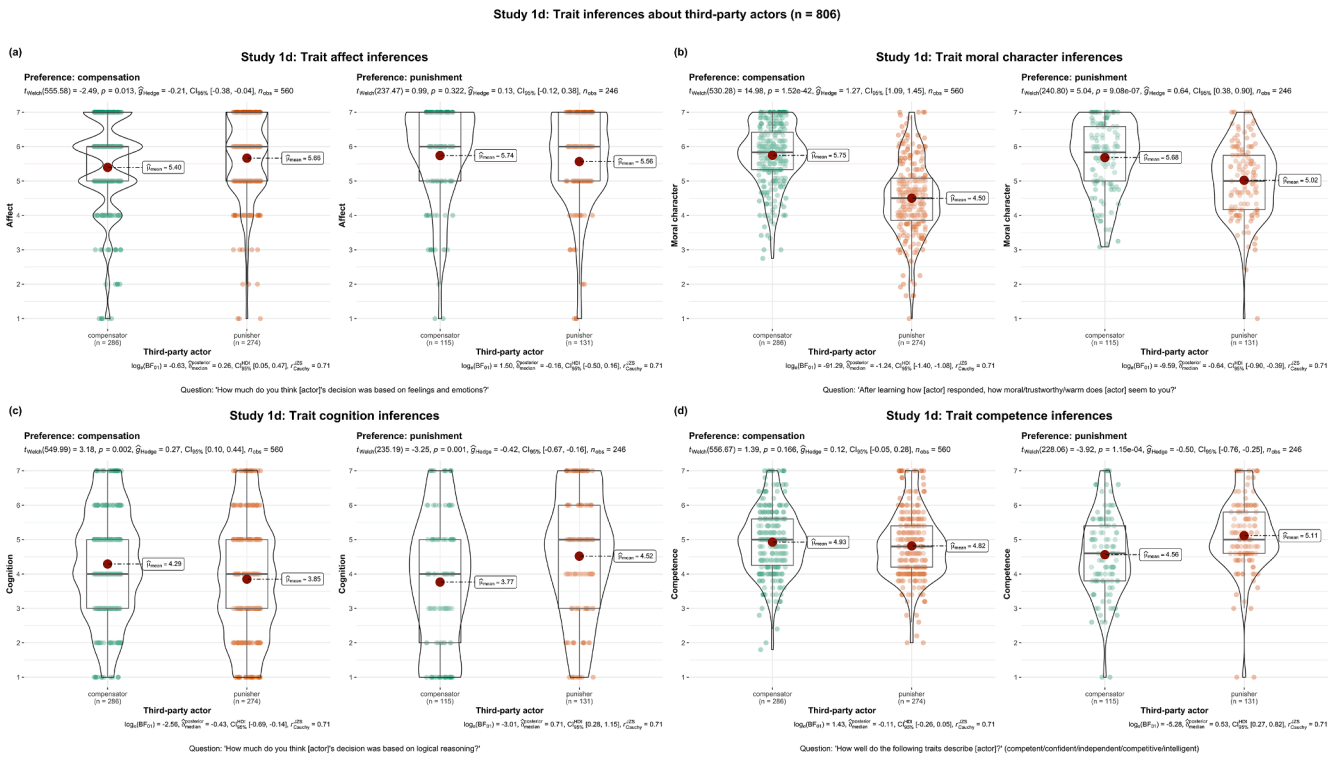

Fig. 3. Third-party choice and inferences about the third party's: (*a*) affect (*b*) moral character (*c*) cognition (*d*) competence, among participants who personally prefer to compensate or punish in response to a third-party norm violation.

no longer receive such reputational benefits. To test this, we include a third-party punishing and compensating game where the amount of money involved is increased by a factor of 10,000.

12. Study 2a:

12.1. Method

To begin, we examined the inferences people make about the trustworthiness and ethicality of those who chose to either (see Fig. 5b): punish, compensate, partly punish *and* partly compensate (henceforth referred to as the "mixed" option), or choose to not intervene (henceforth referred to as the "nothing" option). This was done to determine whether the positive inference that are made about compensators hold when additional options are made available.

A total of 1011 participants participated in this study. These participants were randomly assigned to read that in response to Person A taking from Person B, Person C had chosen to either:

(i) punish A by paying to remove 50 cents from A, or
(ii) compensate B by paying to give 50 cents to B, or
(iii) partly punish A by paying to remove 25 cents from A and partly compensate B by paying to give 25 cents to B, or
(iv) chose to neither punish A nor compensate B.

After reading how C chose to respond, participants responded using a 7-point Likert scale regarding their impressions of how trustworthy and how moral Person C is (averaged to form "character" variable; Pearson's $r = 0.84$). Participants were also asked how they would respond as Person C given the above four options (see Fig. 1b). The order in which these questions were asked was randomized.

12.2. Results

An ANOVA revealed a main effect of condition and an interaction between condition and preference (for full statistical details, see Table 2), such that compensators were perceived to have a better moral character than other actors, but the strength of this effect varied across different personal preferences. Post hoc tests revealed that, collapsing across personal preferences, compensators had significantly higher character ratings than other third-party actors ($p < 0.05$, Bonferroni-corrected; Table 2; see Fig. 5b). The effect size was largest for people who personally preferred compensation (see Supplementary Fig. 2).

13. Study 2b

13.1. Methods

To test whether these reputational benefits for compensators are simply a product of the low-stakes nature of the third-party punishing and compensating game used, we increased the (hypothetical) amount of money involved by a factor of 10,000. A total of 1014 participants took part in this study. The methodology was identical to Study 2a aside from the money at stake, i.e. all monetary units were multiplied by 10,000 (e.g., "Person A took 50¢ from Person B" changed to "Person A took $5000 from Person B").

13.2. Results

The results revealed a main effect of condition, preference, and an interaction between condition and preference (for full statistical details, see Table 2). As can be seen from Fig. 5c, irrespective of whether they personally preferred compensation or punishment, participants continued to rate compensators as more moral and more trustworthy even with the increase in money at stake. The effect size was largest for people who personally preferred compensation (see Supplementary Fig. 3).

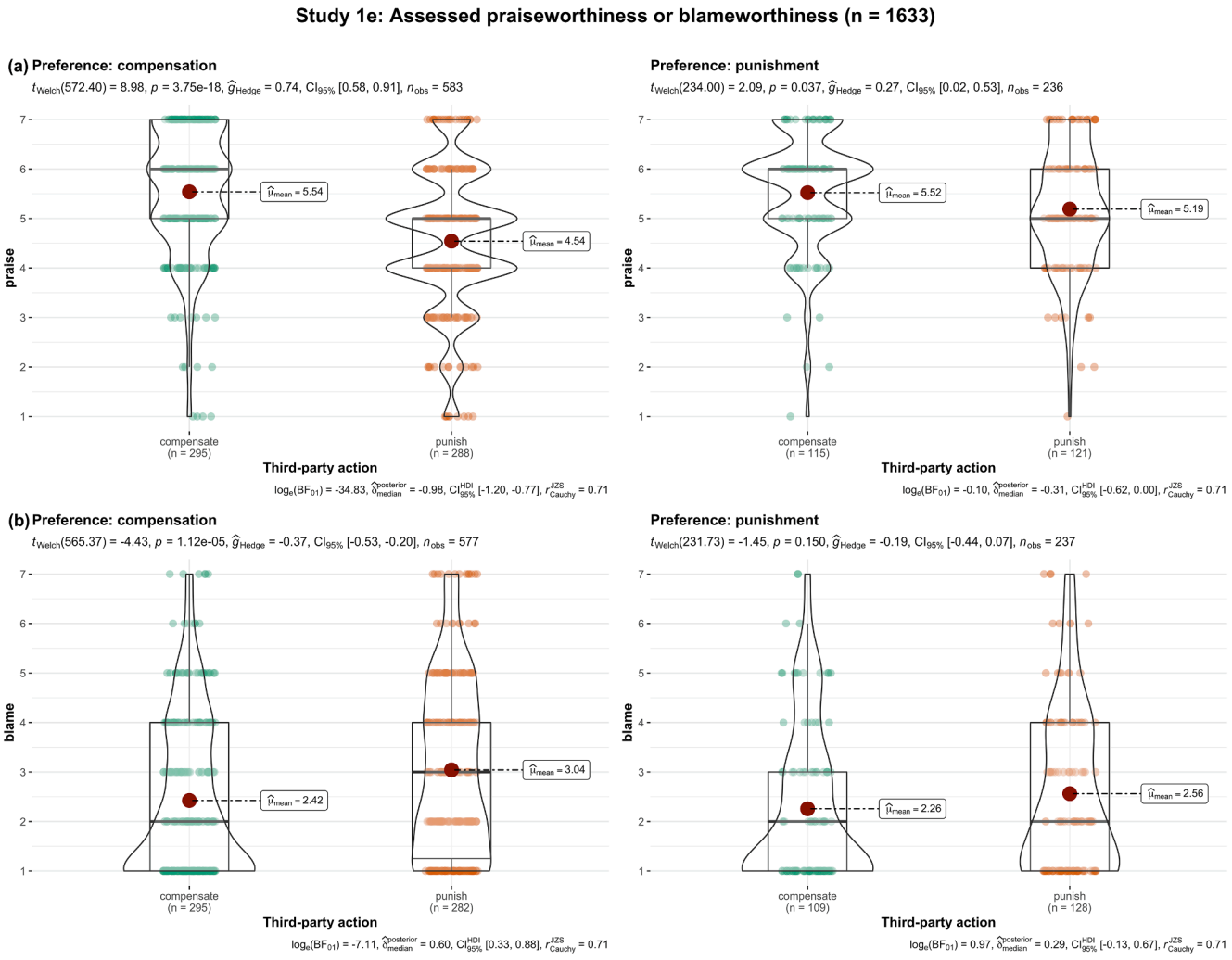

Fig. 4. (*a*) Study 1d found that, irrespective of their personal preference, people found compensation to be the more praiseworthy choice. (*b*) Blame judgments, on the other hand, were dependent on personal preference, specifically, people who preferred compensation found punishers to be more blameworthy.

14. Study 2c

14.1. Methods

Having found that compensators have reputational advantages even in the presence of a broader choice set of third-party responses, we investigated if these reputational benefits scale with the magnitude of the compensation. One additional shortcoming of Study 2a was that the mixed response (part punishment, part compensation) differed from other responses in terms of amount of money the victim received (25 cents, as compared to 50 cents in case of compensation). This was done to have a uniform cost for the third-party intervention across conditions (10 cents), along with uniform "multipliers" translating this cost into a benefit for victims, or punishment for perpetrators. In the current study, we intentionally break this uniformity and change the mixed response from "25 cents removed from transgressor and 25 cents sent to the victim" to "50 cents removed from transgressor and 50 cents sent to the victim". This was done to have a uniform quantity of compensation or punishment across conditions (50 cents).

A total of 2449 participants read instruction for the third-party punishing and compensating game in which Person A and B start with 100 cents, but Person C starts with 20 cents. Participants were then randomly assigned to read that in response to A taking from B, C had chosen one of the following options (see Fig. 1c)-

(i) chose to neither punish A nor compensate B (*nothing*), or
(ii) punish A by paying to remove 50 cents from A (*punisher*), or
(iii) compensate B by paying to give 50 cents to B (*compensator*), or
(iv) partly punish A by paying to remove 50 cents from A and partly compensate B by paying to give 50 cents to B (*mixed*), or
(v) punish A by paying to remove 100 cents from A (*hyperpunisher*), or
(vi) compensate B by paying to give 100 cents to B (*hypercompensator*).

After reading how C chose to respond, participants responded using a 7-point Likert scale regarding their impressions of how trustworthy and how moral is Person C (averaged to form "character" variable; Pearson's $r = 0.81$). Participants were also asked how they would respond as Person C given the above six options. The order in which these questions were asked was randomized.

14.2. Results

Results revealed a main effect of condition, preference, and an interaction between condition and preference (see Table 2; see

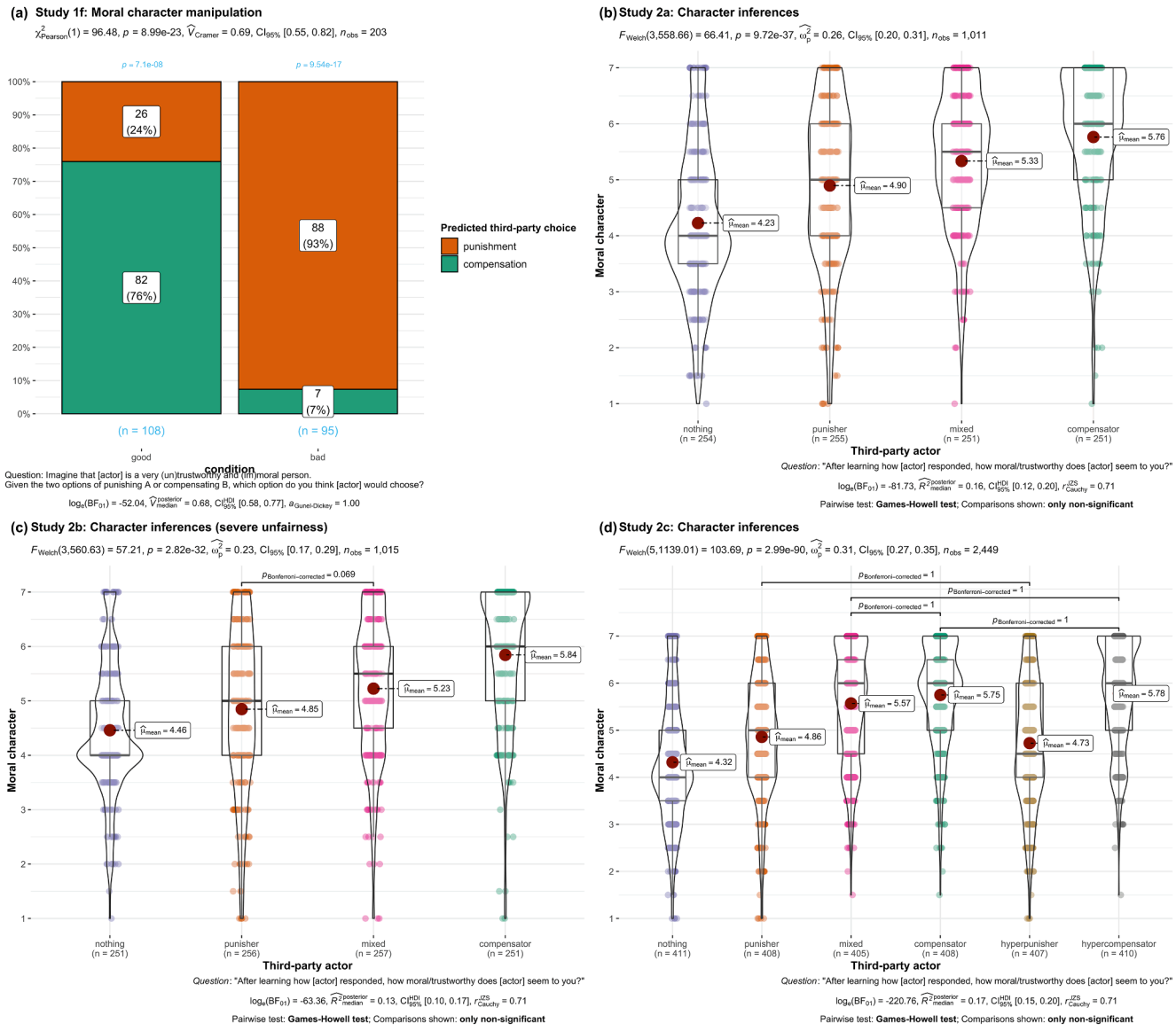

Fig. 5. (*a*) Predictions about how actors of different moral character would choose to respond in response to an observed norm violation. (b) Inferences about the moral character of a third party who chose to either not intervene, punish, punish and compensate, or compensate. (c) Study 2b showed that people inferred third-party compensators to have a better moral character when responding to acts of extreme unfairness. (*d*) Inferences about the moral character of a third party who chose to either not intervene, punish, punish and compensate, compensate, hyperpunish, or hypercompensate. Note that only non-significant comparisons are displayed. That is, all comparisons not shown are statistically significant ($p < 0.05$).

Supplementary Fig. 4 for condition by preference results) such that there were differences in perceived moral character of different actors and these differences were dependent on personal preference of the participant. Given the complexity of this factorial design (6 conditions × 6 preferences = 36 cells), we only focus on post hoc comparisons for the main effect of condition (see Supplementary Fig. 4 for all comparisons). As expected, collapsing across personal preferences, compensators were inferred to have a better moral character than actors in conditions nothing, punishment, and hyperpunishment (see Fig. 5d). Interestingly, hypercompensators were not found to have better moral character than compensators, i.e., compensating the victim more than what they are due doesn't increase inferred ethicality and trustworthiness of the intervening third-party actor. Similarly, actors with hyperpunishment response neither had any advantage or disadvantage compared to their counterparts who engaged only in punishment.

Importantly, actors with mixed response (compensate *and* punish) were found to have equally good moral character as actors who purely compensated. Note that this is a different result from Study 2a where we found that actors with mixed response to have a worse moral character than compensators (Fig. 5b). The difference between these two studies is the *magnitude* of the mixed response. In Study 2a, actors in mixed condition spent 10 cents to remove 25 cents from the perpetrator (who had taken 50 cents) *and* to add 25 cents in compensation to the victim (who had lost 50 cents). In contrast, in Study 2b, actors in mixed condition spent 20 cents to remove 50 cents from the perpetrator (who had taken 50 cents) *and* to add 50 cents in compensation to the victim (who had lost 50 cents). Combining results from Study 2a and 2b, we can assert that to the degree to which compensators *fully* restore victim's losses, it doesn't matter if the compensation is also accompanied by punishment.

15. Part 2 summary: *Additivity of reputational effects*

Although we had established in Part 1 that choosing victim compensation over perpetrator punishment as a third-party actor carries

reputational and cooperative benefits, this was done only in the limited, and rather forced, dichotomous context where the actors could choose only from these two options. In Part 2, we have established that compensators continue to accrue such reputational benefits even with a larger choice set of possible third-party actions (Studies 2a-c). Moreover, across all studies, third parties tended to achieve the ceiling on reputational benefits by compensating the victim just in the amount of her losses. Compensating the victim more than this amount did not provide any additional reputational benefit (Study 2c), nor did engaging in both compensation and punishment. Finally, we found that this effect is not predicated on the magnitude of money involved – compensators are attributed positive moral traits regardless of whether they compensate a loss of 50 cents or $5,000 (Study 2b).

16. Part 3: *Partner choice benefits of compensation*

Having explored the character inferences that are made about punishers and compensators, the next question we investigate is if the character inferences about punishers versus compensators affect subsequent partner choice decisions. Specifically, we gave participants a hypothetical choice between a punisher and cooperator as their future partner in a social economic game—either a Trust Game or a Dictator Game. We reasoned that if the reputational benefits of third-party interventions can be leveraged in cooperative contexts, they should be chosen more frequently to be cooperative partners in economic games requiring trust.

It has been argued that the ultimate adaptive function of making character inferences is to select reliable social partners for cooperative interactions (Baumard et al., 2013). This is an important potential benefit of having a good reputation: by being chosen more often as a social partner, a third party would reap the benefits of social connection thus offsetting the cost of intervening in a third-party role (Barclay & Willer, 2007; Baumard et al., 2013; Fu et al., 2008; Trivers, 1971).

In order to sharpen our focus on this kind of social behavior, often called partner choice, we next built on prior work demonstrating a key signature of it: people choose partners largely based on their underlying social intentions and ignoring accidental variation in the actual outcome of their behavior (Martin & Cushman, 2015). Thus, we tested cases where an actor intends one form of intervention (e.g., victim compensation) but produces an unintended outcome (perpetrator punishment). We predicted that third parties would be chosen as partners for social dilemmas based on their intended response and not based on their actual, but unintended, response. This compliments previous research that highlights the importance of the intentions of the violator (e.g., Desmet, De Cremer, & van Dijk, 2011). We instead highlight the importance of the intentions of the third-party responder. We report the results of the studies comparing intentions versus outcomes in the Supplementary Materials (Studies 3c and 3d).

17. General methods for partner choice studies

Our studies again involved variations of the third-party punishing and compensating game. The underlying structure of the third-party interaction was the same as the one used in Part 1 (Reputational benefits) studies. After reading a description about the third-party interaction, participants were additionally shown, depending on the study, a description of either the Trust Game (TG) or Dictator Game (DG). Past research has shown that cooperative behavior in such economic games reflects a general orientation toward being cooperative, both in other economic games (Peysakhovich, Nowak, & Rand, 2014) as well as non-game measures of cooperation (McAuliffe, Forster, Pedersen, & McCullough, 2018).

In all studies involving these games, there were additional comprehension checks to see if participants understood rules of the game. Participants who failed these checks were still allowed to complete the study, but their data was later discarded during analysis. The basic instructions for each game were as follows:

18. Trust game (TG)

> "In this interaction, you are matched with one other brand-new person. You will be the Proposer and the partner you choose on the next page will be the Decider.
> The Proposer starts with 20 cents.
> First the Proposer makes a choice, then the Decider responds.
> 1) The Proposer can choose to transfer their 20 cents or not.
> If the Proposer transfers 20 cents, then it is TRIPLED and given to the Decider (so
> the decider now has 60 cents).
> 2) The Decider can then choose how many of the cents they want to transfer back to
> to the Proposer (between 0 cents and 60 cents)."

18.1. Dictator game (DG)

> "In this interaction, you are matched with one other brand-new person. You will be the Receiver and the partner you choose on the next page will be the Decider.
> The Decider starts with 100 cents and the Receiver starts with 0 cents.
> This interaction has one single decision:
> 1) The Decider will choose how many of the 100 cents to transfer to the Receiver.
> 2) The Receiver will get the number of cents the Decider transfers and the Decider will get to keep the rest."

If it is true that compensators are perceived to have a better moral character than punishers, this reputational advantage should also lead to cooperative benefits for third-party actors. Specifically, compensators should be preferred as partners in the *Decider* roles in both TG and DG.

19. Study 3a

19.1. Methods

A total of 405 participants were asked how they would respond as Person C. Next, participants were given the instructions for the TG and were asked who they would choose as a partner in the TG: a third-party actor who chose to punish the perpetrator or a third-party actor who chose to compensate the victim.

19.2. Results

Participants more frequently chose the compensator as a partner in the decider role in the TG, but this preference was stronger for people who personally preferred compensation than those who preferred punishment (see Fig. 6a).

20. Study 3b

20.1. Methods

Next, we also explored whether punishers or compensators are chosen more often as deciders for a DG. A total of 405 participants were asked how they would respond as Person C. Participants were then informed of the instructions for the DG and asked who they would choose as a decider in the DG: a third-party punisher or compensator.

20.2. Results

As with the TG, participants chose the compensator more often as the

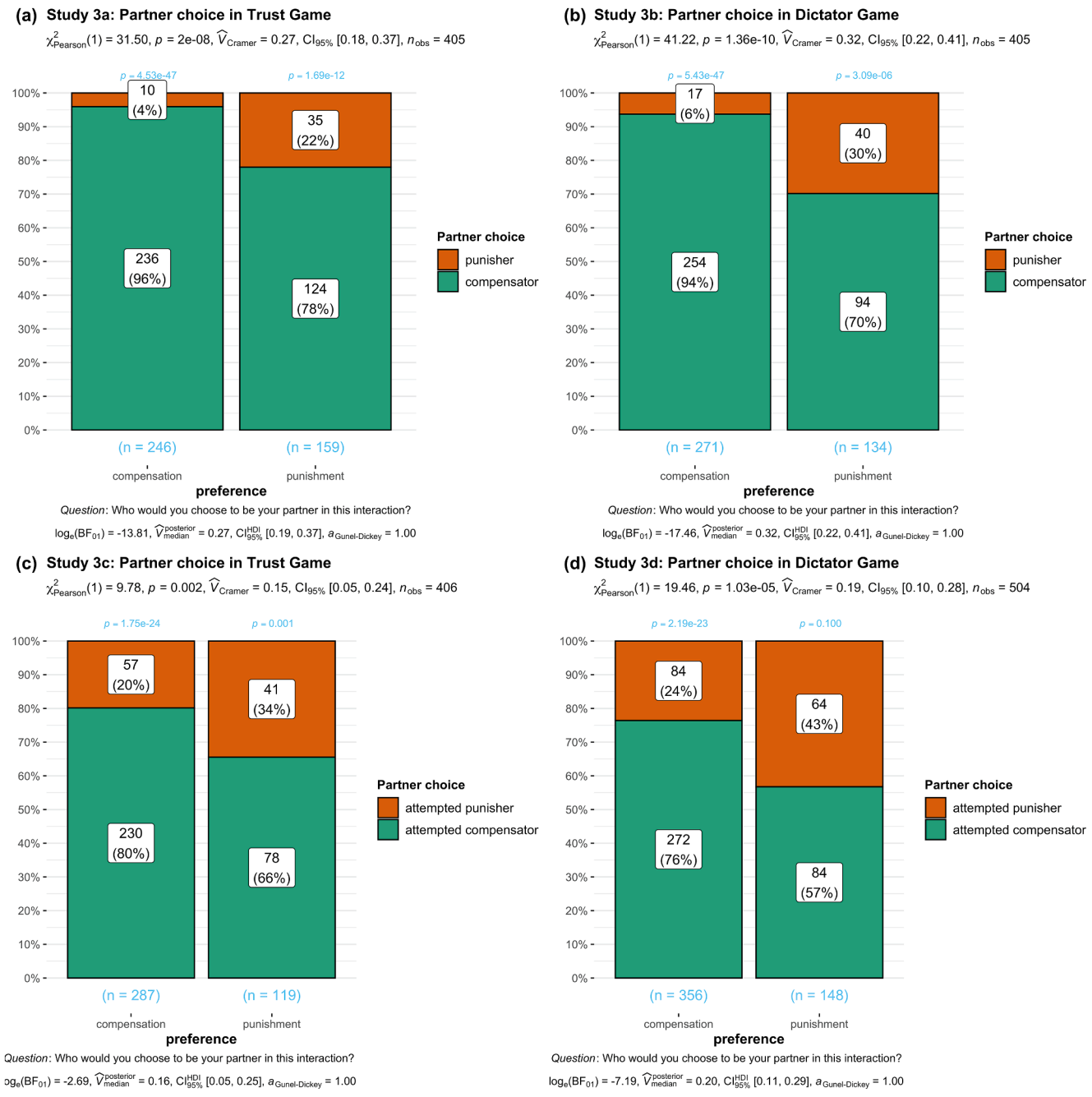

Fig. 6. Choice of a partner: (a) for a Trust Game based on the third party's response (b) for a Dictator Game based on the third party's response (c) for a Trust Game based on the third party's attempted response (d) for a Dictator Game based on the third party's attempted response.

decider for the DG, but this preference was stronger for people who personally preferred compensation than those who preferred punishment (see Fig. 6b).

21. Study 3e

Given the evidence that people prefer compensators as partners, we next explored whether this preference is mediated by their impression of the compensator as a having superior moral character.

21.1. Methods

A total of 935 participants were randomly assigned to evaluate a third party who chose to either punish or compensate in the third-party punishing and compensating game. Next all participants read instructions for the TG and were presented with a choice between the punisher and compensator from the third-party punishing and compensating game and were asked to choose one to be the *Decider* in the TG. Lastly participants were asked whether they would punish or compensate in the third-party punishing and compensating game.

21.2. Results

We carried out robust mediation analysis (implemented in R package *robmed*) with 500 bootstrap samples. As expected from prior studies, regression analyses revealed that-

1. character judgments about third-party actors differed based on condition participants were assigned to (i.e., compensators were perceived to be more moral/trustworthy than punishers) ($B = -1.095$, se $= 0.085$, $z = -12.88$, $p < 0.001$),
2. the better character assessment of third-party actors increased their chances of being selected as cooperative partners ($B = 0.043$, se $= 0.009$, $z = -4.56$, $p < 0.001$),
3. the frequency with which each third-party actor was chosen for the cooperative game differed across conditions ($B = 0.251$, se $= 0.028$, $z = 9.05$, $p < 0.001$),

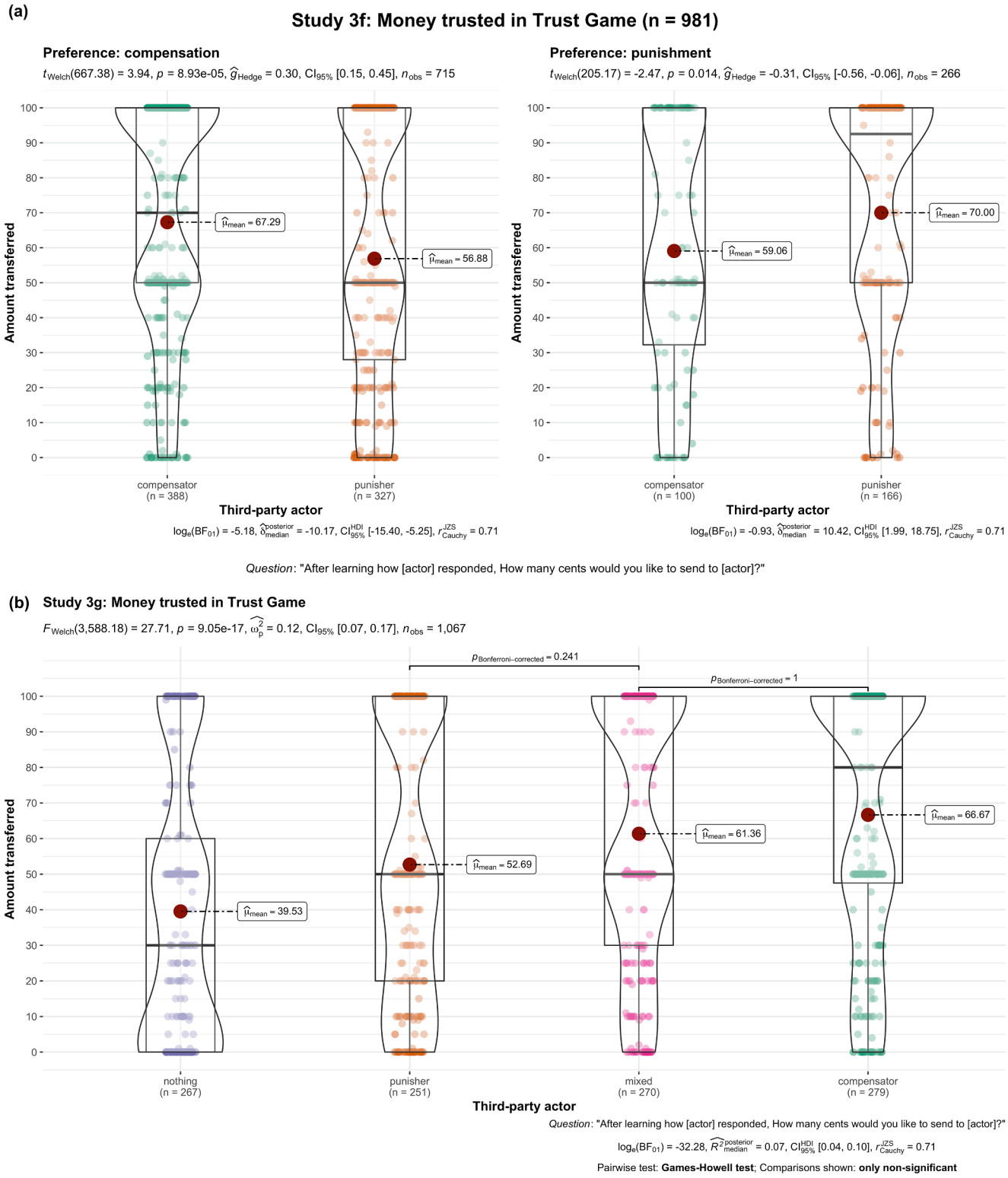

Fig. 7. Amount of money trusted in a Trust Game with a third party who (*a*) chose to either compensate or punish (*b*) chose to not intervene, punish, punish and compensate, or compensate. Note that only non-significant comparisons are displayed. That is, all comparisons not shown are statistically significant ($p < 0.05$).

4. the indirect effect of condition on partner choice was mediated via character judgments ($B = -0.048$, 95% CI_{boot} [−0.071, −0.026]).

22. Study 3f

Studies 3a-e show that when seeking potential cooperative partners, people tend to prefer third-party compensators over third-party punishers. It is natural to think that people would also tend to trust compensators more than punishers during cooperative interactions. Therefore, in the next study we ask: In a trust game, do people tend to trust third-party compensators more than third-party punishers, as measured by the average amount of their (hypothetical) initial endowment "sent" to partners of each type?

22.1. Methods

A total of 981 participants were randomly assigned to read that Person C had chosen either to punish the perpetrator by paying to remove 50 cents, or to compensate the victim by paying to give 50 cents. Next, participants were informed of the instructions for the TG and then were asked how many cents they would like to send to Person C. Then participants were asked how they would respond as Person C.

22.2. Results

In contrast to all prior studies we observed a significant interaction between condition and preference, but no significant main effects of either condition or preference (for full statistical details, see Table 2). In short, participants trusted more money with partners who acted in accordance with their own personal preference. People who personally preferred compensation trusted more money with the compensator, but participants who preferred punishment trusted more money with the punisher (see Fig. 7a).

23. Study 3g

23.1. Methods

Next, we explored whether such trust-related attitudes towards third-party actors was an artifact of the limited choice paradigm by giving the full 4-option version of the game. A total of 1067 participants participated in this study. As in study 3e, participants were randomly assigned to one of the conditions based on who the third-party actor was- nothing, punisher, mixed, and compensator (see Fig. 1b). Like Study 2a, the mixed condition meant paying 5 cents to remove 25 cents from A and paying 5 cents to compensate 25 cents to B. The instructions for the TG were like the previous TG instructions with one notable exception: The Proposer began with 100 cents instead of 20 cents. Participants were told that they would be assigned the role of the Proposer and were asked how many cents they would trust with Person C. Participants were also asked how they would respond as Person C given the above four options. The order in which these questions were asked was randomized.

23.2. Results

Results revealed a main effect of condition, preference, and an interaction between condition and preference (for full statistical details, see Table 2; see Supplementary Fig. 5 for condition by preference results). Collapsing across personal preferences, compensators were trusted with more money than actors who did nothing or chose to punish, but not actors who engaged in a mixed response (see Fig. 7b).

24. Part 3 summary: *Partner choice benefits of compensation*

Consistent with the finding of Part 1 that compensators (vs. punishers) enjoy greater reputational benefits, we found that compensators are more often chosen as cooperative partner in economic interactions that required trust. This feature of partner choice is driven by the intentions of third-party actors to compensate (versus to punish) and not the actual outcomes produced (see Supplementary Results: Studies 3c and 3d).

Unexpectedly, however, a different pattern emerged when we investigated the amounts of resource that people trust to compensators versus punishers (Study 3e and 3f). Here, the most parsimonious interpretation of the data is increased trust for the like-minded: overall, people trusted more money to targets whose third-party intervention behavior matched their own personal preferences. This discrepancy between partner choice, on the one hand, and trust behavior conditioned on assigned partner, on the other, was neither expected nor predicted based on any theoretical framework. That is, we had no *a priori* basis to predict that people would prefer compensators as Trust Game partners but would not exhibit greater trust towards them in the same game. We will not discuss this finding further, but it stands out as an intriguing topic for further study.

Integrating across all our findings, the weight of the evidence suggests important material benefits of third-party compensation (vs. punishment). Compensators enjoy enhanced reputations, they are chosen more often as partners in cooperative interactions, and most (although not *all*) people also put greater trust in them once a cooperative interaction is initiated.

25. Part 4: Are reputational consequences of third-party intervention anticipated by third parties?

Next, we asked whether people successfully anticipate the reputational consequences of third-party interventions. Are they aware of how others will react if they choose to punish, or if they choose to compensate?

Costly signalling theory suggests that adaptations should exist in both the receivers and senders of signals. We have now demonstrated that receivers perceive the signal sent via compensation as one that indicates moral quality. We next examine whether third parties are aware that by compensating rather than punishing they are sending a stronger signal of moral quality. To do so, we explicitly ask participants about the impression they believe others will have of them after they choose to punish or instead to compensate.

In addition, we also ask participants what they believe that *victims* will prefer: A third party who seeks vengeance on their behalf, or one who restores their loss. Previous research shows that people prefer to be compensated as *victims* (Heffner & FeldmanHall, 2019), and prefer to compensate as third parties (van Doorn, Zeelenberg, & Breugelmans, 2018), but choose to enact a combination of both compensation and punishment as third parties if given the option (FeldmanHall, Sokol-Hessner, Van Bavel, & Phelps, 2014; FeldmanHall, Otto, & Phelps, 2018). If people expect that victims have a strong preference for compensation, this may help to explain both why many people choose compensation over punishment in third-party roles, and why many people approve of such choices.

If people are aware that others view compensation more favorably than punishment, they likely represent it as a social norm. And, of course, a wealth of evidence shows that moral choices are influenced by social norms (Gino, Ayal, & Ariely, 2009; Goldstein, Cialdini, & Griskevicius, 2008; Kelly, Ngo, Chituc, Huettel, & Sinnott-Armstrong, 2017; Peysakhovich and Rand, 2016; Salali, Juda, & Henrich, 2015). To the extent that people choose to personally engage in compensation (rather than punishment), this might reflect their knowledge of and conformity to a social norm. Therefore, we also investigate if people's preference for third-party actors is modulated by conformity pressures.

26. Study 4a

26.1. Methods

A total of 204 participants were told that Person A had taken 50 cents from Person B, and participants were assigned the role of Person C (third-party actor, i.e.). After choosing their preferred response, participants were asked to predict how other people informed of that response would rate them on two traits: (1) trustworthiness and (2) morality. Since these measures were highly correlated ($r = 0.80$), they were averaged to give one "moral character" judgment. The order in which these two questions were asked was randomized.

26.2. Results

As expected, participants predicted that compensators would be perceived to have a better moral character as compared to people who preferred punishment (see Fig. 8a; also, see Table 2). Despite this, a substantial number of participants (39% in this study) nevertheless indicated they personally preferred the punitive intervention.

27. Study 4b

27.1. Methods

We next investigated whether participants' personal decisions about

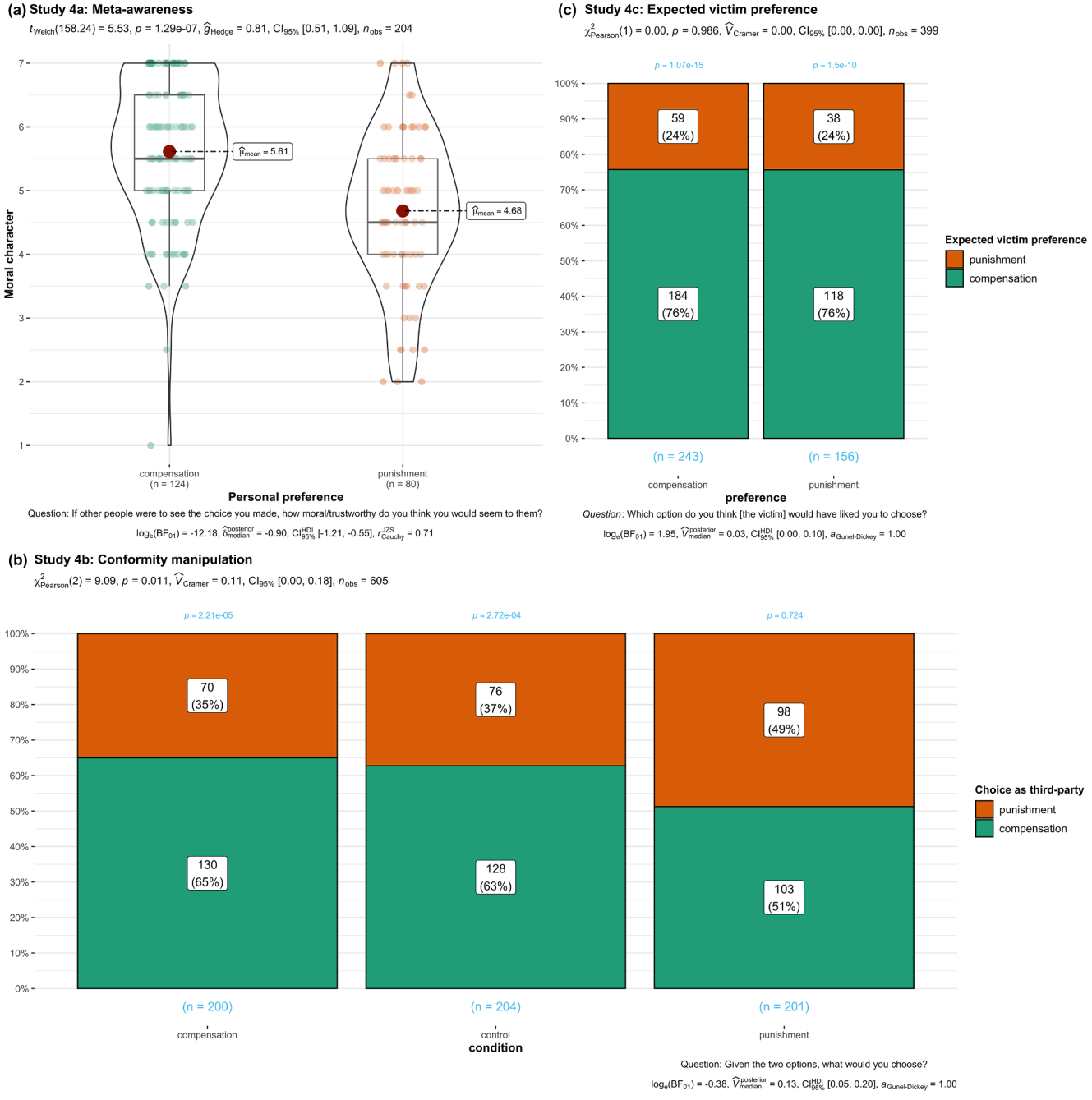

Fig. 8. (*a*) Meta-awareness of inferences drawn by observers about one's moral character based on choice to compensate the victim versus punish the perpetrator. (*b*) Change in participants' preference for action as a third-party after learning that majority of other people chose (left to right): compensation, no information provided, punishment. (*c*) Participants' expectations about what the victims of norm violations would want them to choose.

whether to compensate or punish are sensitive to a direct manipulation of the perceived social norm. A total of 605 participants were told that Person A had taken 50 cents from Person B and were then randomly assigned to one of three conditions:

- the *control* condition: participants were simply asked to choose whether to punish A or compensate B;
- the *punish majority* condition: participants were informed that most previous subjects (percentages varied randomly between 65% and 85%) had chosen to punish and then asked to choose whether to punish A or compensate B;
- the *compensate majority* condition: participants were first informed that most previous subjects (percentages varied randomly between 65% and 85%) had chosen to compensate and then asked to choose whether to punish A or compensate B.

After reading this information, participants gave their choice as Person C (third-party actor, i.e.).

27.2. Results

It should first be noted that this study used deception and therefore concerns may arise about participant suspicion and the resulting pollution of subject pools. While it is worthwhile to avoid using deception when possible, recent evidence suggests that participant level of suspicion is not related to previous experience of deception nor do suspicious participants behave differently compared to non-suspicious participants (Krasnow, Howard, & Eisenbruch, 2019).

Consistent with our prior studies, most participants in the *control* condition chose to compensate rather than to punish ($\chi^2(1) = 12.813$, $p < 0.001$). A similarly sized majority of participants also chose to compensate rather than to punish in the *compensate majority* condition ($\chi^2(1) = 18.182$, $p < 0.001$), when we explicitly instructed participants that most people choose compensation. This preference was eliminated, however, in the *punish majority* condition, when we explicitly instructed participants that most participants choose punishment. Here, participants were equally likely to choose compensation or to choose punishment ($\chi^2(1) = 0.126$, $p = 0.723$) (see Fig. 8b). Taken together, these results suggest that participants are motivated to conform to the statistical norm (explaining why the *punish majority* manipulation influenced their behavior relative to baseline) and, by default, assume that the statistical norm is compensation (explaining why the *compensate majority* manipulation did not influence their behavior relative to baseline). With that said, it may be worth exploring this pattern of results with a focus on more serious/intentional violations given that in such circumstances punishment becomes more normative (Darley & Pittman, 2003).

28. Study 4c

We next examined whether people's expectations about what a victim would prefer (third-party compensation versus third-party punishment) depends upon their own preferences for how to respond as a third party. For instance, do people who personally prefer punishment (or compensation) also expect that victims most desire this response?

28.1. Methods

A total of 400 participants were told that Person A had taken 50 cents from Person B, and as Person C, participants would need to choose whether to punish A or compensate B. After providing their choice of response, participants were asked which option they thought *Person B* (the victim, i.e.) would have liked them to choose.

28.2. Results

A majority (75%) of participants believed that victims would want them to choose compensation over punishment (see Fig. 8c). This basic trend held even among the subset of participants who choose to punish; 76% of such participants predicted that victims would prefer compensation.

29. Part 4 summary: *Anticipated reputational benefits of compensation*

People are aware of the reputational advantage that third-party compensation holds over third-party punishment (Study 4a). They also appear to assume that most people will compensate and therefore only shift their response when descriptive norms favor punishing (Study 4b). Finally, they anticipate that victims would prefer them to engage in compensation (Study 4c). Taken together, these results indicate a detailed and rather accurate representation of the frequency and reputational consequences of compensation versus punishment.

30. Part 5: *Is compensation an honest signal of trustworthiness?*

We have shown in several ways that people tend to think compensators are more trustworthy. We next explore whether this belief is accurate. Do people who engage in third-party compensation, as compared to those who engage in third-party punishment, return more money entrusted to them in a Trust Game?

As discussed before, third-party responding could have become an evolutionarily adaptive behavior because the cost of intervening is offset by the beneficial increase in cooperative opportunities. This requires that social partners cooperate more with third-party interveners, and from an adaptive standpoint this implies that third-party interveners *actually* make for superior cooperative partners. In other words, the signal that is sent via punishing and compensating should turn out to be an honest signal. If such signals are dishonest, people would eventually learn not to trust such signals and third-party responders will lose their adaptive advantage.

Furthermore, our theory is that a particularly strong correlation should exist between the emotions and cognitions for being cooperative and the emotions and cognitions for compensating victims, and this correlation is stronger than the correlation between the emotions and cognitions for being cooperative and the emotions and cognitions for punishing perpetrators. If that is true, we should observe third parties who choose to compensate being more likely to be cooperative with others relative to third parties who choose to punish.

A few recent studies have explored this question within the domain third-party punishing (Jordan, Hoffman, Bloom, & Rand, 2016) as well as within the domain of moral dilemma judgments (Capraro et al., 2018). As it turns out, third-party punishers both appear more trustworthy and *are* more trustworthy than those who choose to not intervene (Jordan, Hoffman, Bloom, & Rand, 2016). Although a few studies have shown association between a certain personality traits and incidental emotions (personality traits: empathic concern, mentalizing, compassion, and moral anger; Civai, Huijsmans, & Sanfey, 2019; Hu, Strang, & Weber, 2015; Leliveld, Dijk, & Beest, 2012; McCall, Steinbeis, Ricard, & Singer, 2014; van Doorn et al., 2018) and a preference for compensating the victim, no study has contrasted third-party compensators' trustworthiness against the trustworthiness of those who choose other intervention strategies.

31. Study 5a

31.1. Methods

A total of 1006 participants were asked how they would respond as Person C after witnessing an interaction where Person A takes from

Person B (see Fig. 9a). Next, participants were informed of the instructions for the TG and were asked to *imagine* that they were the Decider and an anonymous proposer had transferred the entire 20 cents to them. Participants were then asked how many cents they would like to transfer back to this person (from 0 to 60).

31.2. Results

Comparison between two groups revealed that participants who chose to compensate the victim returned more money in the TG compared to participants who chose to punish; i.e., the compensators were indeed more trustworthy than punishers (see Fig. 9a). This was, however, a small effect; the average amount of money returned by compensators was about 5% larger than the average amount returned by punishers, $p = 0.042$.

32. Study 5b

32.1. Methods

In Study 5b, we ran a direct replication of the Study 5a (see Fig. 9b). A total of 950 participants participated in this study. The effect was not significant, but it was again in the direction of compensators being more trustworthy. The effect was again small: In this case, the average amount of money returned by compensators was about 2% larger.

33. Part 5 summary: *Honesty of signals*

We found some evidence that third-party actors who responded to a witnessed norm violation by compensating the victim were slightly more trustworthy than third-party actors who chose to punish the perpetrator. At best, however, our methods detected a small effect. We build upon this finding in Study 6b where we examine whether people who chose to compensate rather than punish in a contextualized vignette study also score lower on the Dark Triad of personality traits (Paulhus & Williams, 2002).

34. Part 6: *Assessing generalizability of the effect*

Until this point, we have focused exclusively on third-party responses to acts of unfairness within an economic game. But fairness norms are just one kind of moral violation (Curry, Chesters, & Van Lissa, 2019; Curry, Mullins, & Whitehouse, 2019). Moreover, we have focused on the behavior of uninvolved third parties whereas, in practice, third party responding is performed within prescribed institutional contexts (Cushman, 2015b). In this next section, we seek to understand whether the general preference for third-party compensators we found extends to other contexts. Moreover, while the third-party punishing and compensating game that was used in the previous studies is a common method to explore the psychology of fairness (e.g. Jordan et al., 2016), it does come with certain shortcomings. First, the stakes are low given that one player is only stealing cents from another player (although we explored hypothetical judgments of high stakes violations Study 2b).

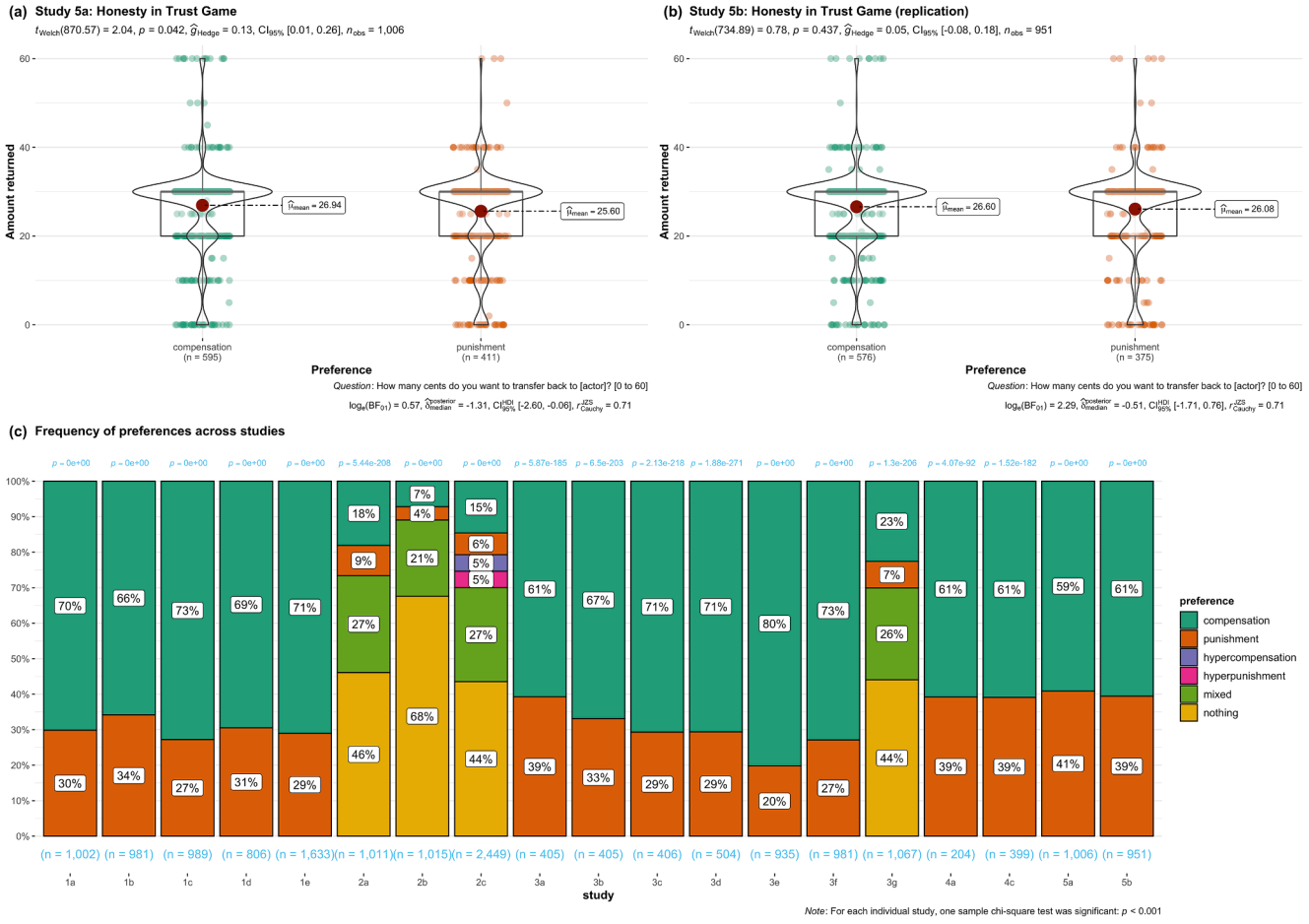

Fig. 9. (a) Amount of money third-party compensators and punishers transferred back in a Trust Game. (b) Replication of Study 5a. (c) Frequency of different personal preferences across studies. When given the dichotomous choice of either choosing compensation or punishment, most people personally preferred compensation. When a more thorough choice set was provided, the most frequently chosen option was to do nothing.

Second, it is rather abstract in that people are only given superficial details about the players involved. Third, when Player A steals from Player B, Player A is not breaking any rules of the game (since the experimenters have stipulated that this is allowed in the game). Fourth, the money for compensation comes from the third-party whereas in many contexts the money might come from other sources such as tax revenue. Our next studies examine whether our findings persist when we vary all such factors.

In an initial pilot study along these lines we described a person who encounters one person violently battering another in a park and either punishes (apprehending the perpetrator) or compensates (coming to the aid of the victim). This study returned equivocal results, showing that neither punishment nor compensation was strongly preferred. This prompted us to conduct a more systematic investigation. We report those results across Studies 6a, 6b, and 6c.

35. Study 6a

We explored the generality of our findings across a variety of different kinds of moral violations: (*a*) the Bernie Madoff scandal where a politician could choose to either prosecute and sentence Madoff, or instead to establish a victim compensation fund, (*b*) bike theft on a university campus where the Student Body President could choose to either use limited available funds to work with campus police to catch thieves, or instead to give temporary bikes and free locks to the victims of the theft, (*c*) the Darfur crisis where a politician could form a committee to focus on either punishing the perpetrators, or instead on compensating the victims, (*d*) domestic violence case, where a city counsellor could choose to either spend grant money to hire more police officers and increase the attorney general's budget to prosecute offenders, or instead to increase the budget to provide counselling and financial aid to victims, (*e*) verbal abuse case, where an uninvolved third-party could intervene and either admonish the perpetrator, or instead console the victim. The bike theft, Darfur crisis, and domestic violence scenarios were taken from previous research (Adams & Mullens, 2013).

Together, these five scenarios range from contexts where the third-party is an institutional actor whose job it is to address moral violations to contexts where the third-party is uninvolved and does not have an obligation to address the moral violation. Also, some of these violations were more likely to evoke concern about unfairness or, in any event, possession (the Madoff scandal, the bike theft violation) and some were more likely to evoke concern about physical or emotional harm (the Darfur crisis, domestic violence, verbal abuse).

35.1. Methods

This study was a within-subjects design where 435 participants (after excluding those who failed to pass manipulation checks) were presented with each of the five vignettes in random order (see Supplementary Text for full vignettes) and were asked to respond to 2–4 comprehension checks depending on the vignette. Participants were then told after each vignette whether the third-party chose to compensate or punish. After reading how the third-party chose to respond, participants responded using a 7-point Likert scale regarding their impressions of how trustworthy and how moral the third party is (averaged to form "character" variable; Pearson's $r = 0.78$). Participants were also asked how they would respond as the third party to each scenario given the two options of compensate or punish.

35.2. Results

Linear mixed-effects analysis revealed that indeed, across the diverse set of conditions we utilized, compensators were perceived to have a

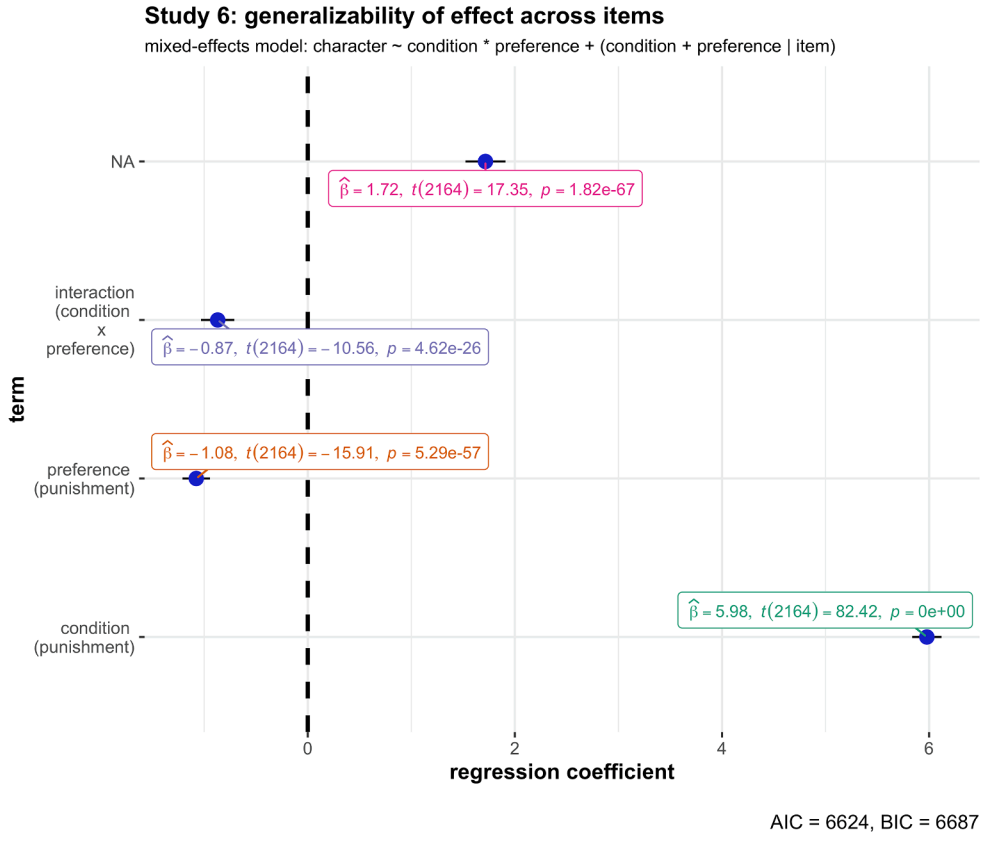

Fig. 10. Linear mixed-effects analysis shows that the preference for third-party compensators over punishers (as assessed by moral/trustworthiness ratings) is generalizable across multiple different types of moral violation contexts.

better moral character than punishers (see Fig. 10; also see Supplementary Fig. 6), although the strength of the effect varied based on the personal preferences. Therefore, it is unlikely that the effects we have observed thus far were due to the specific situation/context we have been working with.

36. Study 6b

In Study 6a we find that our pattern of results generalizes across a wide variety of moral violations. In Study 6b we utilize a workplace vignette where we ask participants to imagine being an employee at an organization where a moral violation has occurred and where the leader chooses to either punish, compensate, or engage in both responses (see Supplementary Text for full vignettes). By doing so, we examine whether compensators who fail to remove the violator from the workplace are still preferred by those who may be at risk of being targeted by that same violator. Said differently, we further examine whether compensators are preferred even when punishing would bring greater collective benefit to the group.

We also test a central aspect of our theoretical framework, namely whether people infer that punishers may be riskier cooperation partners. Specifically, we test whether people see punishers as harbouring greater psychopathic tendencies relative to compensators. We also examine in this study whether our finding from Study 2c regarding the mixed response replicates in a contextualized workplace vignette. That is, are workplace leaders who engage in a mixed response (compensate *and* punish) seen as equal, and not superior, moral actors relative to those who only engage in compensation. We also build upon our previous studies by examining whether punishers are seen as better suited for certain leadership roles and whether people shift their preference from compensating the victim to punishing the perpetrator depending on certain reputational incentives. Lastly, we examine whether participants who report preferring to punish the violator rather than compensate the victim in our workplace vignette score higher on the Dark Triad of personality traits (Paulhus & Williams, 2002): Machiavellianism (Christie & Geis, 1970), narcissism, and psychopathy (Jones & Paulhus, 2014). This further examines our preliminary evidence of honest signalling in Part 5.

36.1. Methods

This study was a between-subjects design where 644 participants (after excluding those who failed to pass the comprehension questions) were presented with a vignette describing a workplace scenario and were asked to imagine being an employee at the organization. In the vignette, another employee was wrongfully refused a promotion three years prior and now the injustice has come to light (see Supplementary Text for full vignettes). The leader of the organization chooses to either fire the perpetrator (a manager) who refused to give the promotion, promote and compensate the victim, or engage in both responses (mixed response).

Participants were then asked to provide their impression of the leader of the organization. Participants responded using a 9-point Likert scale regarding their impressions of the third party's trustworthiness and morality. Participants also provided their impression, using a 7-point Likert scale, of the leader's benevolence, integrity, and competence (Mayer, Davis, & Schoorman, 1995), their desire to work for the leader and be loyal to the leader, as well their impression of the leader's psychopathic traits (Jones & Paulhus, 2014). Participants also assessed how well suited the leader was for the role of a *i*) tough, competent leader versus *ii*) warm, caring leader. Next all participants were asked if they were the leader whether they would choose to punish or compensate. All participants were also asked whether they would choose to punish or compensate if they were wanting to develop a reputation of being a *i*) tough, competent leader and a *ii*) warm, caring leader. Lastly, participants completed a measure of the Dark Triad of personality traits Paulhus & Williams, 2002): Machiavellianism (Christie & Geis, 1970), narcissism, and psychopathy (Jones & Paulhus, 2014). Prior to conducting this study, methods, hypotheses, and analysis plans were preregistered and can be accessed at: https://aspredicted.org/blind.php?x=ku63zj.

36.2. Results

Compared to punishers, both compensators and mixed responders were viewed as more trustworthy, moral, benevolent, and competent (for full statistical details, see Table 3). They were also seen as having more integrity and participants reported a greater desire to work for them and be loyal to them. There was no difference between compensators and mixed responders on such measures. On the other hand, punishers and mixed responders were seen as equally psychopathic whereas compensators were seen as uniquely less psychopathic. Participants also showed a strong preference to compensate the victim (compensate = 74% vs. punish = 26%), X^2 (1, N = 644) = 149.22, p < 0.001, and continued to show that preference when the incentives favored appearing as a warm, caring leader (compensate = 78% vs. punish = 22%), X^2 (1, N = 644) = 201.24, p < 0.001, but that preference was reversed when the incentives favored appearing as a tough, competent leader (compensate = 39% vs. punish = 61%), X^2 (1, N = 644) = 29.57, p < 0.001. Participants also viewed punishers and mixed responders as better suited than compensators to be tough, competent leaders and compensators to be better suited than both punishers and mixed responders to be warm, caring leaders. Lastly, we found that participants who chose to compensate rather than punish scored lower on measures of trait Machiavellianism (M = 3.44, SD = 0.74 vs. M = 3.78, SD = 0.42) t(642) = 5.61, p < 0.001, narcissism (M = 3.63, SD = 1.14 vs. M = 4.31, SD = 0.75) t(642) = 7.15, p < 0.001, and psychopathy (M = 3.08, SD = 1.31 vs. M = 4.40, SD = 1.02) t(642) = 11.83, p < 0.001.

37. Study 6c

In Study 6b participants read about a workplace violation and evaluated the third party who addressed the situation by either punishing or compensating. Here we build upon Study 6b by examining people's actual lived experience of witnessing or hearing about a third party addressing a moral conflict.

Table 3
Character evaluation of punisher, compensator, and mixed responder from Study 6b.

Judgment of Trait	Punisher	Compensator	Mixed Responder	F	p
Trustworthy	M = 5.80[a]	M = 7.50[b]	M = 7.65[b]	78.92	<0.001
	SD = 2.20	SD = 1.36	SD = 1.36		
Moral	M = 5.74[a]	M = 7.50[b]	M = 7.59[b]	77.50	<0.001
	SD = 2.20	SD = 1.45	SD = 1.43		
Benevolent	M = 4.52[a]	M = 5.77[b]	M = 5.75[b]	80.12	<0.001
	SD = 1.50	SD = 0.97	SD = 0.96		
Integrity	M = 4.58[a]	M = 5.66[b]	M = 5.73[b]	69.18	<0.001
	SD = 1.45	SD = 0.95	SD = 0.92		
Competent	M = 5.06[a]	M = 5.74[b]	M = 5.68[b]	27.77	<0.001
	SD = 1.22	SD = 0.95	SD = 0.93		
Work with	M = 4.57[a]	M = 5.96[b]	M = 5.90[b]	74.81	<0.001
	SD = 1.71	SD = 1.04	SD = 1.11		
Loyal to	M = 4.47[a]	M = 5.64[b]	M = 5.77[b]	60.09	<0.001
	SD = 1.69	SD = 1.16	SD = 1.10		
Psychopathic	M = 4.10[a]	M = 3.34[b]	M = 3.92[a]	10.78	<0.001
	SD = 1.57	SD = 1.97	SD = 1.80		
Suited to be Tough Leader	M = 5.36[a]	M = 4.91[b]	M = 5.54[a]	13.70	<0.001
	SD = 1.19	SD = 1.48	SD = 1.16		
Suited to be Warm Leader	M = 4.31[a]	M = 6.07[b]	M = 5.74[c]	96.09	<0.001
	SD = 1.81	SD = 1.01	SD = 1.24		

Note: Means with different superscripts differ (p < 0.05).

37.1. Methods

This study was a between-subjects design where 396 participants (after excluding those who failed to pass the comprehension questions) were randomly assigned to either: a) write about an experience from their own life where they witnessed or heard about a third party standing up to a violator for harming a victim or b) write about an experience from their own life where they witnessed or heard about a third party supporting a victim of a harmful violation. After writing about this experience, participants were then asked to provide, using a 9-point Likert scale, their impressions of the third party's trustworthiness and morality. Participants also provided their impression, using a 7-point Likert scale, of the third party's benevolence and integrity (Mayer, Davis, & Schoorman, 1995) and psychopathic traits (Jones & Paulhus, 2014). Lastly, participants were asked on a 7-point Likert scale how severe the violation was and how long ago the violation occurred. Prior to conducting this study, methods, hypotheses, and analysis plans were pre-registered and can be accessed at: https://aspredicted.org/blind.php?x=2v6cf6.

37.2. Results

Compared to participants who wrote about a third party who stood up to a violator who harmed a victim (punishment condition), participants who wrote about a third party who supported a victim of a harmful violation (compensation condition) tended to evaluate the third party as more trustworthy, moral, benevolent, as having marginally more integrity, and as less psychopathic (for full statistical details, see Table 4). Compared to participants who wrote about a third party who stood up to a violator, participants who wrote about a third party who supported a victim evaluated their chosen violation as being more serious. The majority (73%) of participants who wrote about a third party who stood up to a violator, as well as the majority (74%) of participants who wrote about a third party who supported a victim of a harmful violation, wrote about a violation that occurred less than five years ago.

38. Part 6 summary: *Assessing generalizability of the effect*

In this section, we found that the general preference for third-party compensators extends to contexts of harmful, high stakes, concrete, norm violations where either designated or non-designated third parties occupy the role of punisher/compensator and where the cost to intervene varied widely. Furthermore, within the workplace context (Study 6b), a mixed response did not garner any greater reputational benefits relative to a pure compensatory response. Thus, although mixed responders are providing the greatest benefit to the group (helping the victim while removing the bad actor), people do not see them as significantly more trustworthy than pure compensators. This may in part stem from the fact that by engaging in punishment the mixed responder is increasing the perception that they harbour greater psychopathic tendencies relative to pure compensators, which we document in Study 6b. We also find in Study 6b that people shift their choice between punishing and compensating depending on the reputational incentives present in their environment, and importantly, we find that compensating is a robust, honest signal of having lower levels of immoral personality traits. Finally, we find in Study 6c that our pattern of results replicates when examining people's actual lived experience of witnessing third parties confronting violators and supporting victims.

Table 4
Character evaluation of punishers versus compensators from Study 6c.

Judgment of Trait	Punisher	Compensator	t	p
Trustworthy	$M = 7.38$ $SD = 1.68$	$M = 7.91$ $SD = 1.41$	3.40	=0.001
Moral	$M = 7.69$ $SD = 1.67$	$M = 8.07$ $SD = 1.36$	2.49	=0.01
Benevolent	$M = 5.78$ $SD = 1.13$	$M = 6.11$ $SD = 0.98$	3.12	=0.002
Integrity	$M = 5.77$ $SD = 1.12$	$M = 5.97$ $SD = 1.05$	1.82	=0.07
Psychopathic	$M = 2.68$ $SD = 1.40$	$M = 2.08$ $SD = 1.26$	4.47	<0.001
Seriousness of Violation	$M = 5.48$ $SD = 1.39$	$M = 5.90$ $SD = 1.27$	3.14	=0.002

39. Part 7: *Frequency of preferences across study*

Finally, we analyze the frequency of personal preferences that people exhibited across all of our studies and found that, when studies included the binary choice of either compensating or punishing, participants consistently preferred compensation over punishment. This aligns with previous findings showing that third parties prefer to compensate rather than punish when forced to choose between the two options (van Doorn, Zeelenberg, & Breugelmans, 2018). This is interesting because although compensating is psychologically rewarding (Curry et al. 2018), neurobiological research suggests that people find punishing more psychologically rewarding than compensating (Stallen et al., 2018). When the additional options of choosing a mixed response (partly punish, partly compensate) or choosing non-intervention were provided, participant showed the strongest preference for not intervening in the situation, followed by a preference for the mixed response. This finding is in tension with prior research that found that non-intervention was a relatively unpopular choice (FeldmanHall et al., 2014). It also relates to debates regarding the supposed bystander effect and whether people have a desire to intervene to help others (Latané & Darley, 1970; Latané & Nida, 1981; Philpot et al., 2019). Finally, it raises the possibility that the usual laboratory practice of providing dichotomous choice set to participants may not tap into their real preferences (Balafoutas, Nikiforakis, & Rockenbach, 2014; Pedersen, McAuliffe, & McCullough, 2018).

When the additional options of hyperpunishment and hypercompensation were provided, only a small minority (9.26% total) chose these options, revealing that people are not interested either in compensating the victims more than what they are due or punishing the perpetrator in a manner incommensurate with transgression severity (see Fig. 9c for results from Parts 1–5). Future research could explore whether compensators or punishers experience different levels of decision conflict while making their decision about how to intervene (Mata, 2019), whether these preferences for compensators vary depending upon whether a third party is responding to a loss or a gain that is being distributed unfairly (Liu, Li, Zheng, & Guo, 2017), and how the reputation of those who compensate victims and punish perpetrators compares to the reputation of those who reward helpers (de Kwaadsteniet, Kiyonari, Molenmaker, & van Dijk, 2019).

40. General discussion

We conducted a comprehensive examination of the reputational and cooperative benefits that are afforded by two forms of third-party intervention: Punishment and compensation. Although punishment offers clear reputational benefits compared to doing nothing, compensation was still the most favored strategy by the observers. Across various contexts ranging from economic games (Parts 1–5) to workplace injustice (Study 6b) to people's actual personal lives (Study 6c), we find that compensating victims leads to greater reputational and partner choice benefits relative to punishing perpetrators. It is not that punishers are disliked, but rather people seem to have a particular preference for compensators. In the current work, third-party compensators were perceived to be more trustworthy and were chosen more frequently as partners in cooperative interactions that required trust, although they were not (always) trusted with more money. These trends tended to hold even among people who would personally have chosen third-party

punishment. In sum, within the context examined, third-party compensation stands out as a superior reputation enhancement mechanism to third-party punishment.

We began by examining the reputational benefits of these two forms of third-party intervention. Within our economic game, the robust reputational benefits to compensators tended to be specific to personality traits relevant to cooperation. For instance, compensators were viewed by observers as having a superior moral character, irrespective of the observer's stated preference to intervene by punishment versus compensation themselves. In contrast, other kinds of character inferences that are less directly relevant to cooperation were contingent upon people's personal preferences for punishment versus compensation. For example, participants who personally preferred to punish found compensators to be more moral but found punishers to be more competent.

These findings bear on current theories of trait inference and person perception. First, they add further qualification to recent work suggesting that people who are judged to be more moral are also judged to be more competent (Stellar & Willer, 2018). It may be that when judging outright immoral acts there is a convergence between judgments of competence and morality. But, when comparing two altruistic acts, our economic game results suggest that judgments of competence and morality may diverge. With that said, our workplace vignette results did show a convergence between ratings of morality and competence with compensators and mixed responders being rated as more moral and more competent than punishers. But when asked to choose which leader would be better suited to be a tough, competent leader, more people selected the punisher rather than the compensator. This comparison between the reputation of punishers and compensators can be seen as contributing to research on the character dimensions of agency/competence and communion/warmth (Abele, Cuddy, Judd, & Yzerbyt, 2008; Abele & Wojciszke, 2007; Digman, 1997; Fiske, 2012; Fiske, Cuddy, & Glick, 2007; Judd, James-Hawkins, Yzerbyt, & Kashima, 2005; Wiggins, 1991). Second, these results add further qualification to recent work showing that people who rely on emotion (vs. reason) when cooperating are expected to cooperate more (Levine et al., 2018). We found that people who themselves prefer to compensate saw punishers as being more emotional and less logical, and ultimately less trustworthy. This suggests that a perceived reliance on emotion, even when engaging in altruistic acts, can signal inferior character traits.

Returning to our main results, we next examined participant judgments about the *act* of punishing and compensating. Here, we found that compensating was seen as more praiseworthy than punishing, irrespective of participant's personal preference. People also predict that good people are more likely to choose to compensate rather than to punish. Taken together, these results suggest a consistency when examining third-party intervening from both an act-based as well as a person-centered approach (Landy and Uhlmann, 2018; Robinson et al., 2017; Uhlmann, Pizarro, & Diermeier, 2015).

We next examined whether positive trait inferences about third parties translated into the benefit of being chosen as cooperation partners (Barclay & Willer, 2007; Baumard et al., 2013; Fu et al., 2008; Trivers, 1971). We found that compensators were indeed chosen more often than punishers as partners for the Trust Game as well as the Dictator Game. Additionally, in line with prior work that has argued for supremacy of the intent information for partner choices (Martin & Cushman, 2015), we found that participants cooperated with actors with the mere *intent* to compensate the victim, even when unrealized.

Having established that choosing to compensate carries both reputational and cooperative benefits over choosing to punish, we investigated if this result was a methodological artifact of the forced choice between punishment and compensation, which is not representative of real-world decision making. We found that compensators continued to accrue substantial reputational benefits even with a larger choice set of possible third-party actions. Notably, compensating the victim beyond the magnitude of their loss (i.e. hypercompensating) doesn't provide any additional reputational benefits, a finding that converges with recent work suggesting that subtle prosocial acts can signal more trustworthiness than grand prosocial displays (Bird, Ready, & Power, 2018). Additionally, we found across both our economic game and workplace vignette (Study 6b) that once a third party had compensated a victim fully for their losses, there was little evidence of an incremental reputational value to additionally punishing the perpetrator. In contrast, having punished, there was incremental value to additionally compensating.

Next, we examined adaptive design in the psychology of the signaller and found that participants anticipated that choosing compensation would lead others to infer that they (participants) have a superior moral character. This aligns with recent work on meta-awareness when making decisions in moral dilemmas which found that people can anticipate the kind of inferences observers make after observing their actions and such social considerations causally contribute to their decision-making (Rom & Conway, 2018). We also found evidence that people engage in the most prevalent form of third-party intervention in the general population—i.e., to adhere to descriptive norms. Also, participants believed that the victims would prefer compensation over the perpetrator being punished within our economic game.

We also examined using both the TG and psychometric measures of moral personality traits whether the signal sent via third-party compensation is an honest signal. When using the TG, we found inconclusive evidence of whether those who choose to compensate in our economic game were more trustworthy. On the other hand, when using psychometric measures of moral personality traits, we found robust evidence that those who choose to compensate the victim in our workplace vignette were also less Machiavellian, narcissistic, and psychopathic than those who choose to punish the perpetrator. This aligns with our theoretical framework that a particularly strong correlation should exist between the emotions and cognitions associated with compensating and with cooperating.

Furthermore, while compensators are perceived to be more moral than punishers, these reputational benefits might not always translate into a preferential assessment of compensators. Our findings illustrate some downsides to third-party compensating. For example, we find in Study 1c that among individuals who personally prefer the punishment intervention, compensators are viewed as less logical and less competent. We also find in Study 6b that punishers are seen as better suited for the role of a tough, competent leader and participants choose to punish rather than compensate when the reputational incentives favour appearing tough and competent. It may be the case that people who choose to punish understand that while compensators may seem more moral, punishment is the more effective means of promoting good behavior and therefore is the more logical response. That is, people can perceive compensation to be a good *act* and a compensator to be a good *person*, but still might not think that compensation is the right *choice*. Indeed, we find that for a few of our behavioral measures, people's own personal preference for how to respond to a moral violation moderated evaluation of other people's third-party responses. For example, people who personally prefer punishment attributed better moral traits to compensators and chose them as cooperators more frequently but (in Studies 3f and 3 g) exhibited greater trust in third-party punishers by endowing them with more money in the Trust Game. Although personality trait evaluation conditional on personal preferences is a new avenue for signaling research, we hazard a few possible explanations as to why individuals might differ in terms of their assessment of punishers versus compensators.

Another concern that might motivate negative evaluation of compensators (versus punishers) for individuals with a preference for punishment is that this will lead to a reduced likelihood that one's group will abide by moral norms. Additionally, although compensation recuperates victim's losses, it may leave them vulnerable to future exploitation by the same or other perpetrators (McNulty, 2011). While punishment may not garner an individual the same type of reputational benefits, it is an

effective way to promote cooperation within a group (Balliet et al., 2011; Fehr & Gächter, 2000; Mathew & Boyd, 2011) and deters future exploitation (Krasnow et al., 2012; Krasnow, Delton, Cosmides, & Tooby, 2016). In an attempt to examine the boundaries of our effect we examine in Part 6 whether third-party compensators are still preferred when responding to severe violations. We found that compensators are stilled preferred within such contexts. Furthermore, we find this preference is even seen when people are asked to recall from their own lives instances of witnessing a third party who either confronted a violator or supported victims. Thus, it appears that our findings may not simply be a product of the hypothetical nature of many of our experiments. With that said, future research is still needed to determine the boundary conditions of our effect.

41. Theoretical Framework: Costly signaling

It is apparent why compensating a victim would tend to enhance a person's reputation: It is a prosocial action designed to remediate an antisocial one. What is less apparent is why, as suggested by previous work (Jordan, Hoffman, Bloom, & Rand, 2016) and confirmed by our own, the reputational benefit of compensation exceeds that of punishment. Our approach to this question is grounded in costly signaling theory.

A key insight of costly signaling theory is that it can only be successful when there is an underlying, independent relationship between individual differences in cost and individual differences in the trait to be signaled (Gintis, Smith, & Bowles, 2001; Grafen, 1990; Roberts, 1998; Spence, 1973; Zahavi, 1975). Thus, to use costly victim compensation to signal future prosociality, it must be the case that those who pay less of a cost to compensate victims are also more likely to have the requisite means or interest in future prosocial behavior. Likewise for punishment. This is our point of entry.

Prior research has applied these principles to explain why those who engage in third-party punishing are seen as more trustworthy than those who do not respond (Jordan, Hoffman, Bloom, & Rand, 2016). It posits that cooperative individuals find it less costly than non-cooperative individuals to punish because there likely exists a correlation between the incentives for being cooperative and the incentives for punishing immoral behaviour. For instance, if an individual profits from the relationship with a friend, then they will find it less costly to cooperate with that friend and also less costly to defend that friend's interests by punishing harm doers: The material costs of these actions are partially offset by the benefits of promoting a valuable friendship.

This suggests two possible explanations for why compensation would enhance a person's reputation more than punishment. First, to the extent that the incentives for cooperating with a person are correlated with the incentives for punishing on their behalf, presumably they are even more closely correlated with the incentives for compensating them. Like ordinary acts of cooperation, compensation involves a direct transfer of benefits to its recipient, whereas punishment provides a more unique and indirect form of benefit (i.e., by deterring the perpetrator from harming the victim again). Thus, if cooperating with a person has benefits, it seems relatively more likely that compensating them for harm would have comparable benefits, and relatively less likely that punishing on their behalf would have comparable benefits.

Second, and relatedly, there are salient incentives for engaging in punishment other than to benefit the victim. First, punishment can be used to deter the transgressor (or others) from harming the punisher, rather than to deter harm to the victim or other third parties (Delton & Krasnow, 2017; Krasnow, Delton, Cosmides, & Tooby, 2016). Second, punishment can be motivated by the desire to impose a direct cost on its target for purely competitive reason, independent of their transgression (Herrmann, Thöni, & Gächter, 2008). Sometimes, for instance, a person exploits the social "excuse" afforded by punishment in order to thwart a rival's interests (Herrmann, Thöni, & Gächter, 2008). To the extent that a person is motivated by these concerns, it would diminish the correlation between the incentives for cooperation and punishment upon which successful costly signaling depends.

Apart from costly signaling theory, we also consider the possibility that compensation enhances one's reputation more than punishment because its immediate effect is to help, rather than to harm, a person. Notwithstanding that punishment is often held to be a morally justified form of harm, people may still harbor reservations about the character of a person who engages in it. We know that punishment is often triggered by the experience of moral outrage and anger (Darley & Pittman 2003) while third-party compensation may be more likely the result of cooperation-relevant emotions such as empathic concern, mentalizing, and compassion (Civai, Huijsmans, & Sanfey, 2019; Darley & Pittman 2003; Hu, Strang, & Weber, 2015; Leliveld, Dijk, & Beest, 2012; McCall, Steinbeis, Ricard, & Singer, 2014). While empathy and compassion are typically perceived as exclusively positive traits (cf. Bloom, 2017), recent experimental evidence suggests people view high levels of anger as a signal of negative character traits (Gaertig, Barasch, Levine, & Schweitzer, 2019).

42. Why punish?

We and others find that third-party compensation provides greater reputational benefits than third-party punishment. Yet, reputation has been advanced as a principle explanation for third party punishment. Given that greater reputational benefits can be obtained by compensating, why do third parties sometimes punish?

One possible route through which third-party punishment evolved is through higher-order punishment, i.e. punishing those who fail to punish (Cinyabuguma, Page, & Putterman, 2006; Fu, Ji, Kamei, & Putterman, 2017; Kiyonari & Barclay, 2008; Martin, Jordan, Rand, & Cushman, 2019). A second possibility is that while compensators are seen as more moral, by punishing, a third party can be seen as more formidable and as a result be less likely to be the target of harm in the future (Krasnow et al., 2012; Krasnow, Delton, Cosmides, & Tooby, 2016). This may be especially important in cultures of honor where having a reputation of being formidable is particularly beneficial (Nisbett & Cohen, 1996). A third possibility is that punishment allows punishers to equalize or elevate payoffs relative to perpetrators, and therefore provides punishers with benefits as a direct result of decreasing the perpetrator's relative welfare (Raihani & Bshary, 2019). A fourth possibility is that punishment serves as a sufficient moral reputation enhancing mechanism only when the option for compensation is not available to the third party (Jordan & Rand, 2019). A fifth possibility is that while an individual can benefit more from compensating, a group benefits more by having people who are willing to punish and enforce moral norms (Fehr and Fischbacher, 2003; Henrich et al., 2006; Henrich et al., 2010). This can only occur, however, if the between group selection pressures are greater than the individual benefits of not punishing, which may be unlikely (Boyd, Gintis, Bowles, & Richerson, 2003; Henrich, 2004; Henrich & Boyd, 2001; Sober & Wilson, 1999). Determining the relative influence of such mechanisms would serve a valuable avenue for future research to explore.

43. Practical implications

Our research also provides valuable psychological insights into the growing interest in victim compensation as an alternative to violator punishment (Chouhy, 2018). Recent nationally representative research on white-collar crime has shown that people support victim compensation that is paid for not only by the offender, but also a third party, such as taxpayer funded compensation from the government (Galvin, Loughran, Simpson, & Cohen, 2018). We hope this research serves as a catalyst for future inquiry into this area. For example, research is needed to assess perceptions of compensators in other contexts where compensation is seen as particularly less costly than punishment. For example, punishers may be preferred when compensation is seen as

relatively cheap way to avoid punishing the violator and implementing justice. In such contexts, would compensators still enjoy greater reputational benefits? Particularly when the violator has committed a heinous crime? Of course, in many serious contexts such as murder, compensation is not truly possible, thus punishment is seen as form of compensation. How would an attempt at compensation appear in such a context? Furthermore, in many contexts the distinction between punishing and compensating is not clear cut—for instance standing between the violator and victim or calling the police could be seen as both confronting the violator and helping the victim. Our Study 6c examined this nuance but future research is needed to explore the multiple variation between pure punishment and pure compensation. With that said, our findings outline some potential consequences for third parties who choose to move away from punishment and toward compensation as a means of addressing moral violations.

44. Limitations

Several limitations of our study should be addressed in future work. Although we demonstrate that compensators gain greater reputational benefits than punishers across numerous contexts, the majority of our research questions were examined using economic games. This lack of context ideally aids in the generalizability of our findings, but future research is needed to determine whether each of our findings replicate across different contexts and across non-MTurk samples. This is particularly the case given that previous research has shown that in response to specific types of violations, victims in fact prefer that the perpetrator be punished rather than they themselves be compensated (Reb et al., 2006). For instance, within the context of particularly heinous crimes, perhaps compensation alone would be seen as insufficient, particularly if the compensation is seen as an easy and cheap way to avoid implementing justice. Future research is thus needed to explore our numerous research questions across other violation types and contexts.

CRediT authorship contribution statement

Nathan A. Dhaliwal: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Indrajeet Patil:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - review & editing, Visualization. **Fiery Cushman:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.obhdp.2021.01.003.

References

Abele, A. E., Cuddy, A. J., Judd, C. M., & Yzerbyt, V. Y. (2008). Fundamental dimensions of social judgment. *European Journal of Social Psychology, 38*(7), 1063–1065.

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*(5), 751.

Adams, G. S., & Mullen, E. (2013). Increased voting for candidates who compensate victims rather than punish offenders. *Social Justice Research, 26*(2), 168–192.

Andersen, S., Ertaç, S., Gneezy, U., Hoffman, M., & List, J. A. (2011). Stakes matter in ultimatum games. *American Economic Review, 101*(7), 3427–3439.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences, 111*(45), 15924–15927.

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137*(4), 594–615.

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

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society of London B: Biological Sciences, 274*(1610), 749–753.

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences, 36*(1), 59–78.

Bird, R. B., Ready, E., & Power, E. A. (2018). The social significance of subtle signals. *Nature Human Behavior, 1*. https://doi.org/10.1038/s41562-018-0298-3.

Bloom, P. (2017). *Against empathy: The case for rational compassion*. Random House.

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science, 328*(5978), 617–620.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*(6), 3531–3535.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13*(3), 171–195.

Boyd, R., & Richerson, P. J. (2005). *The origin and evolution of cultures*. Oxford University Press.

Brambilla, M., & Leach, C. W. (2014). On the importance of being moral: The distinctive role of morality in social judgment. *Social Cognition, 32*(4), 397–408.

Capraro, V., Sippel, J., Zhao, B., Hornischer, L., Savary, M., Terzopoulou, Z., … Griffioen, S. F. (2018). People making deontological judgments in the Trapdoor dilemma are perceived to be more prosocial in economic games than they actually are. *PloS One, 13*(10), Article e0205066.

Christie, R., & Geis, F. (1970). *Studies in Machiavellianism*. New York: Academic Press.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics, 9*(3), 265–279.

Civai, C., Huijsmans, I., & Sanfey, A. G. (2019). Neurocognitive mechanisms of reactions to second- and third-party justice violations. *Scientific Reports, 9*, 9271. https://doi.org/10.1038/s41598-019-45725-8.

Chouhy, C. (2018). Moving beyond punitive interventions: Public support for government-funded victim compensation for white-collar crime victims. *Criminology & Public Policy, 17*(3), 547–551.

Curry, O. S., Chesters, M. J., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality, 78*, 106–124.

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate. *Current Anthropology, 60*(1), 47–69.

Curry, O. S., Rowland, L. A., Van Lissa, C. J., Zlotowitz, S., McAlaney, J., & Whitehouse, H. (2018). Happy to Help? A systematic review and meta-analysis of the effects of performing acts of kindness on the well-being of the actor. *Journal of Experimental Social Psychology, 76*, 320–329.

Cushman, F. (2015a). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology, 6*, 97–103.

Cushman, F. (2015b). Punishment in humans: From intuitions to institutions. *Philosophy Compass, 10*(2), 117–133.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review, 7*(4), 324–336.

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior, 38*(6), 734–743.

de Kwaadsteniet, E. W., Kiyonari, T., Molenmaker, W. E., & van Dijk, E. (2019). Do people prefer leaders who enforce norms? Reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology, 84*, Article 103800.

Desmet, P. T., De Cremer, D., & van Dijk, E. (2011). In money we trust? The use of financial compensations to repair trust in the aftermath of distributive harm. *Organizational Behavior and Human Decision Processes, 114*(2), 75–86.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 73*(6), 1246.

Eisenbruch, A. B., & Roney, J. R. (2017). The skillful and the stingy: Partner choice decisions and fairness intuitions suggest human adaptation for a biological market of cooperators. *Evolutionary Psychological Science, 3*(4), 364–378.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology, 71*(2), 390.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review, 90*(4), 980–994.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature, 422*(6928), 137.

FeldmanHall, O., Otto, A. R., & Phelps, E. A. (2018). Learning moral values: Another's desire to punish enhances one's own punitive behavior. *Journal of Experimental Psychology: General, 147*(8), 1211–1224. https://doi.org/10.1037/xge0000405.

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.

Fiske, S. T. (2012). Warmth and competence: Stereotype content issues for clinicians and researchers. *Canadian Psychology/Psychologie Canadienne, 53*(1), 14.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83.

Fu, F., Hauert, C., Nowak, M. A., & Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Physical Review E, 78*(2), Article 026117.

Fu, T., Ji, Y., Kamei, K., & Putterman, L. (2017). Punishment can support cooperation even when punishable. *Economics Letters, 154*, 84–87.

Gaertig, C., Barasch, A., Levine, E. E., & Schweitzer, M. E. (2019). When does anger boost status? *Journal of Experimental Social Psychology, 85*, Article 103876.

Galvin, M. A., Loughran, T. A., Simpson, S. S., & Cohen, M. A. (2018). Victim compensation policy and white-collar crime: Public preferences in a national willingness-to-pay survey. *Criminology & Public Policy, 17*(3), 553–594.

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science, 20*(3), 393–398.
Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology, 213*(1), 103–119.
Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research, 35*(3), 472–482.
Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*(1), 148.
Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology, 144*(4), 517–546.
Heffner, J., & FeldmanHall, O. (2019). Why we don't always punish: Preference for non-punitive responses to moral violations. *Scientific Reports, 9*, 13219. https://doi.org/10.1038/s41598-019-49680-2.
Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization, 53*(1), 3–35.
Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.
Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208*(1), 79–89.
Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., … Lesorogol, C. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science, 327*(5972), 1480–1484.
Henrich, N., & Henrich, J. P. (2007). *Why humans cooperate: A cultural and evolutionary explanation*. Oxford University Press.
Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Lesorogol, C. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767–1770.
Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362–1367.
Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience, 9*, 24. https://doi.org/10.3389/fnbeh.2015.00024.
Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment, 21*(1), 28–41.
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.
Jordan, J. J., & Rand, D. G. (2019). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology, 118*(1), 57.
Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology, 89*(6), 899.
Kelly, M., Ngo, L., Chituc, V., Huettel, S., & Sinnott-Armstrong, W. (2017). Moral conformity in online interactions: Rational justifications increase influence of peer opinions on moral judgments. *Social Influence*, 1–12.
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.
Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What are punishment and reputation for? *PLOS One, 7*(9), Article 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.
Krasnow, M. M., Howard, R. M., & Eisenbruch, A. B. (2019). The importance of being honest? Evidence that deception may not pollute social science subject pools after all. *Behavior Research Methods*, 1–14.
Landy, J. F., & Uhlmann, E. L. (2018). Morality is personal. In K. Gray, & J. Graham (Eds.), *The atlas of moral psychology: Mapping good and evil in the mind* (pp. 121–132). New York: Guilford.
Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* New York: Appleton-Century-Crofts.
Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin, 89*(2), 308–324. https://doi.org/10.1037/0033-2909.89.2.308.
Leliveld, M. C., Dijk, E., & Beest, I. (2012). Punishing and compensating others at your own expense: The role of empathic concern on reactions to distributive injustice. *European Journal of Social Psychology, 42*(2), 135–140.
Levine, E. E., Barasch, A., Rand, D., Berman, J. Z., & Small, D. A. (2018). Signaling emotion and reason in cooperation. *Journal of Experimental Psychology: General, 147*(5), 702.
Li, J., Li, S., Wang, P., Liu, X., Zhu, C., Niu, X., … Yin, X. (2018). Fourth-party evaluation of third-party pro-social help and punishment: An ERP study. *Frontiers in Psychology, 9*, 932.
Liu, Y., Li, L., Zheng, L., & Guo, X. (2017). Punish the perpetrator or compensate the victim? Gain vs. Loss context modulate third-party altruistic behaviors. *Frontiers in Psychology, 8*, 2066.
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), Article e0125193. https://doi.org/10.1371/journal.pone.0125193.
Martin, J. W., Jordan, J. J., Rand, D. G., & Cushman, F. (2019). When do we punish people who don't? *Cognition, 193*, Article 104040.
Mata, A. (2019). Social metacognition in moral judgment: Decisional conflict promotes perspective taking. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspa0000170.
Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences, 108*(28), 11375–11380.
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.
McAuliffe, W., Forster, D., Pedersen, E. J., & McCullough, M. (2018). Does cooperation reflect the operation of a broad trait? *European Journal of Personality*.
McCall, C., Steinbeis, N., Ricard, M., & Singer, T. (2014). Compassion meditators show less anger, less punishment, and more compensation of victims in response to fairness violations. *Frontiers in Behavioral Neuroscience, 8*, 424.
McNulty, J. K. (2011). The dark side of forgiveness: The tendency to forgive predicts continued psychological and physical aggression in marriage. *Personality and Social Psychology Bulletin, 37*(6), 770–783.
Nakamaru, M., & Iwasa, Y. (2006). The coevolution of altruism and punishment: Role of the selfish punisher. *Journal of Theoretical Biology, 240*(3), 475–488.
Nelissen, R. M. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior, 29*(4), 242–248.
Nisbett, R. E., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the South*. Boulder, CO, US: Westview Press.
O'Gorman, R., Wilson, D. S., & Miller, R. R. (2005). Altruistic punishing and helping differ in sensitivity to relatedness, friendship, and future interactions. *Evolution and Human Behavior, 26*(5), 375–387.
Ohtsuki, H., Iwasa, Y., & Nowak, M. A. (2009). Indirect reciprocity provides a narrow margin of efficiency for costly punishment. *Nature, 457*(7225), 79.
Patil, I. (2018). ggstatsplot: "ggplot2" Based Plots with Statistical Details. CRAN. Retrieved from https://cran.r-project.org/web/packages/ggstatsplot/index.html.
Patil, I., Zanon, M., Novembre, G., Zangrando, N., Chittaro, L., & Silani, G. (2017). Neuroanatomical basis of concern-based altruism in virtual environment. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2017.02.015.
Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience, 9*(1), 94–107.
Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36*(6), 556–563.
Pedersen, E. J., McAuliffe, W., & McCullough, M. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. doi: 10.1037/xge0000410.
Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Communications, 5*, 4939. https://doi.org/10.1038/ncomms5939.
Peysakhovich, A., & Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science, 62*(3), 631–647.
Philpot, R., Liebst, L. S., Levine, M., Bernasco, W., & Lindegaard, M. R. (2019). Would I be helped? Cross-national CCTV footage shows that intervention is the norm in public conflicts. *American Psychologist*.
Raihani, N. J., & Bshary, R. (2015a). The reputation of punishers. *Trends in Ecology & Evolution, 30*(2), 98–103.
Raihani, N. J., & Bshary, R. (2015b). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution, 69*(4), 993–1003.
Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences, 1*.
Rand, D. G., Armao, J. J., IV, Nakamaru, M., & Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology, 265*(4), 624–632.
Reb, J., Goldman, B. M., Kray, L. J., & Cropanzano, R. (2006). Different wrongs, different remedies? Reactions to organizational remedies after procedural and interactional injustice. *Personnel Psychology, 59*(1), 31–64.
Richerson, P. J., & Boyd, R. (1998). The evolution of human ultrasociality. *Indoctrinability, Ideology, and Warfare: Evolutionary Perspectives*, 71–95.
Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago: University of Chicago Press.
Roberts, G. (1998). Competitive altruism: From reciprocity to the handicap principle. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 265*(1394), 427–431.
Robinson, J. S., Page-Gould, E., & Plaks, J. E. (2017). I appreciate your effort: Asymmetric effects of actors' exertion on observers' consequentialist versus deontological judgments. *Journal of Experimental Social Psychology, 73*, 50–64.
Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology, 74*, 24–37.
Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology, 69*, 44–58.
Salali, G. D., Juda, M., & Henrich, J. (2015). Transmission and development of costly punishment in children. *Evolution and Human Behavior, 36*(2), 86–94.
Smith, E. A., & Bird, R. L. B. (2000). Turtle hunting and tombstone opening: Public generosity as costly signaling. *Evolution and Human Behavior, 21*(4), 245–261.
Sober, E., & Wilson, D. S. (1999). *Unto others: The evolution and psychology of unselfish behavior, No. 218*. Harvard University Press.
Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics, 87*(3), 355–374.
Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C. K., & Sanfey, A. G. (2018). Neurobiological mechanisms of responding to injustice. *Journal of Neuroscience, 38*(12), 2944–2954.

Stellar, J. E., & Willer, R. (2018). Unethical and inept? The influence of moral information on perceptions of competence. *Journal of Personality and Social Psychology, 114*(2), 195.
Teper, R., Tullett, A. M., Page-Gould, E., & Inzlicht, M. (2015). Errors in moral forecasting: Perceptions of affect shape the gap between moral behaviors and moral forecasts. *Personality and Social Psychology Bulletin, 41*(7), 887–900.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.
Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72–81.
Van Doorn, J., Zeelenberg, M., & Breugelmans, S. M. (2018). An exploration of third parties' preference for compensation over punishment: Six experimental demonstrations. *Theory and Decision, 85*(3–4), 333–351.
Van Doorn, J., Zeelenberg, M., Breugelmans, S. M., Berger, S., & Okimoto, T. G. (2018). Prosocial consequences of third-party anger. *Theory and Decision, 84*(4), 585–599.
Wiggins, J. S. (1991). Agency and communion as conceptual coor- dinates for the understanding and measurement of interpersonal behavior. In W. Grove, & D. Cicchetti (Eds.), *Personality and Psychopathology: Vol. 2. Thinking clearly about psychology: Essays in honor of Paul E. Meehl* (pp. 89–113). Minneapolis: University of Minnesota Press.
Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology, 53*(1), 205–214.