Acta Psychologica 205 (2020) 103060

Contents lists available at ScienceDirect

Acta Psychologica

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

Distinct affective responses to second- and third-party norm violations

Lauren E.S. Hartsough*, Matthew R. Ginther[1], René Marois

Department of Psychology, Vanderbilt University, PMB 407817, 2301 Vanderbilt Place, Nashville, TN, 37240, United States of America

ARTICLE INFO

Keywords:
Punishment
Second-party
Third-party
Emotions
Moral outrage

ABSTRACT

Social norm violations provoke strong emotional reactions that often culminate in punishment of the wrongdoer. This is true not only when we are the victims of the norm violation (second-party), but also when witnessing a complete stranger being victimized (third-party). What remains unclear, however, is whether second- and third-party punishments are associated with different emotions. To address this question, here we examine how subjects respond affectively to both second- and third-party norm violations in an economic game. Our results indicate that while second- and third-parties respond to norm violations by punishing wrongdoers similarly, they report experiencing distinct emotional states as a result of the violation. Specifically, we observed a cross-over interaction between anger and moral outrage depending on the party's context: while anger was more frequently reported for second- than for third-party violations, moral outrage was more evoked by third-party than second-party violations. Disgust and sadness were the most prevalently reported emotions, but their prevalence were unaffected by party contexts. These results indicate that while responses to second- and third-party violations result in similar punishment, they are associated with the expression of distinct affective palettes. Further, our results provide additional evidence that moral outrage is a critical experience in the evaluation of third-party wrongdoings.

1. Introduction

Emotions shape behavior. They do so not only by motivating and regulating our own behavior, but also by affecting the behavior of those surrounding the emotive person (Plutchik, 1980). This two-pronged effect of emotion on behavior is particularly salient for the punishment of norm violators, a behavior thought to have been critical for the evolution of societal cooperation and a key element of our modern criminal justice system (Bowles & Gintis, 2011; Fehr & Fischbacher, 2004). However, while we have strong visceral responses to norm violations and these responses predict punishment, the specific mechanisms and emotions involved remain poorly understood (Carlsmith, Darley, & Robinson, 2002; Treadway et al., 2014). Even less well understood is whether the same affective responses drive the two main ways norm violators are punished: second-party punishment (2PP), wherein the victim punishes the offender themselves; and third-party punishment (3PP), wherein uninvolved individuals punish the offender. The two behaviors have largely been studied independently of one another, resulting in two tracts of literature exploring second- and third-party emotional responses to norm violations. Each has yielded distinct results, as described below.

Second-party responses to norm violations are primarily studied using economic games (ex. De Quervain et al., 2004; Eriksson, Strimling, Andersson, & Lindholm, 2016). Increasingly unfair behaviors evoke greater amounts of 2PP, driven by both fairness concerns and a desire for revenge (Bone & Raihani, 2015). Further, increased amounts of 2PP in response to unfair behavior such as free-riding is largely motivated by negative emotions and leads to increased cooperation (Fehr & Gachter, 2002). Anger, particularly, has been identified as the primary emotional response to unequal treatment affecting oneself, and is associated with increased 2PP (Russell & Giner-Sorolla, 2013; Small & Loewenstein, 2005). Consistent with this finding, anger, together with disgust and contempt, was recently found to motivate a desire to punish others in response to real-life immoral acts against oneself (Hofmann, Brandt, Wisneski, Rockenbach, & Skitka, 2018).

Third-party responses to norm violations have been explored using hypothetical scenarios as well as economic games (ex. Landmann & Hess, 2016; Lotz, Okimoto, Schlosser, & Fetchenhauer, 2011). Studies of the underlying emotions initially focused predominantly on contempt, anger, and disgust—building on the 'CAD Triad' hypothesis, which first hypothesized these emotions as the primary responses to third-party norm violations (Rozin, Lowery, Imada, & Haidt, 1999;

* Corresponding author.
E-mail address: lauren.hartsough@vanderbilt.edu (L.E.S. Hartsough).
[1] Present Address: United States District Court for the District of Massachusetts, 1 Courthouse Way, Boston, MA 02210, United States of America.

https://doi.org/10.1016/j.actpsy.2020.103060
Received 15 April 2019; Received in revised form 4 March 2020; Accepted 9 March 2020

Shweder, Much, Mahapatra, & Park, 1997). However, these studies did not ultimately support the CAD model: Contempt is evoked mainly in response to violations attributed to incompetence and does not appear to predict 3PP (Hutcherson & Gross, 2011). Expressions of anger and disgust tend to be highly correlated to one another in response to norm violations and predict 3PP similarly, leading some to believe that disgust is merely an anger-synonym and not a distinct response to third-party violations (Nabi, 2002; Russell, Piazza, & Giner-Sorolla, 2013).

While contempt, anger, and disgust have received the lion's share of 3PP research, recent work has called attention to the role of an emotional state known as 'moral outrage' in driving 3PP (Carlsmith et al., 2002; Lotz et al., 2011; Salerno & Peter-Hagene, 2013). Moral outrage has been poorly characterized in the literature, with many defining it simply as a negative affective response to perceived norm violations that drives 3PP (Carlsmith, Darley, & Robinson, 2002; Darley & Pittman, 2003). Salerno and Peter-Hagene (2013) suggested that moral outrage is a combination of anger and disgust, though they did not compare these emotions to the construct of moral outrage directly and their measure of moral outrage conflated moral outrage with a desire to punish. It has been hypothesized that moral outrage is specific to third-party norm violations while anger is evoked in response to violations against oneself or a cared-for victim (Batson, Kennedy, & Nord, 2007; Landmann & Hess, 2016) though this hypothesis has not yet been empirically tested. It appears to be distinct from disgust and contempt as moral outrage (like anger) typically motivates "approach" behaviors, such as direct punishment, while disgust and contempt may lead to avoidance behaviors, such as shunning (Carver & Harmon-Jones, 2009; Molho, Tybur, Guler, Balliet, & Hofmann, 2017; Van de Vyver & Abrams, 2015). In a recent study we assessed the emotional responses and punishment decisions of subjects in response to criminal scenarios that varied in both the severity of the harm to the victim and in the intent of the perpetrator. We found that moral outrage was the emotion that was expressed most strongly in response to severe harm with culpable intent, and that it was moral outrage – not contempt, anger, or disgust – that mediated the effects of the interaction between intent and harm in driving 3PP decisions (Ginther, Hartsough, & Marois, under review). While these findings establish moral outrage as a key emotional response mediating 3PP, they could not establish whether moral outrage is uniquely expressed in 3PP or whether it also governs all forms of norm-based punishment, including 2PP.

While it would be parsimonious and reasonable to conclude that the same affective state underlies both 2PP and 3PP, the studies cited above might suggest otherwise. Correspondingly, there is evidence suggesting that these two behaviors are cognitively and evolutionarily distinct (Gummerum & Chu, 2014; Riedl, Jensen, Call, & Tomasello, 2012; Strobel et al., 2011). Studies comparing second- and third-party emotional responses suggest that anger tends to be evoked when oneself is the target of a norm violation, while disgust is evoked more when another person is the target (Hutcherson & Gross, 2011; Molho et al., 2017). It has been suggested that disgust and contempt may relate to judgment of character while anger is sensitive to the self-relevance of the norm violation (Hutcherson & Gross, 2011). Batson et al. (2007) proposed that anger is experienced when the interests of oneself or a cared-for other have been thwarted, while moral outrage is evoked when a moral standard has been violated even when it does not impact oneself (e.g. unfairness towards a third-party). They found that participants expressed anger in response to unfair behavior towards themselves or someone they had been primed to feel empathy towards, but did not express anger in response to unfairness against another whom they were not primed to empathize with. The authors interpreted this as evidence against the existence of moral outrage. However, the unfair behavior in this study was a selfish allocation of tickets to maximize potential payout, which some have suggested does not actually represent a norm violation as the individual is simply acting in their best interest (Dunning, Anderson, Schlosser, Ehlebracht, & Fetchenhauer, 2014). This may account for studies that have found limited emotional responses and punishment in third-party contexts, as many relied on unequal outcomes without an explicit norm violation. In addition to this limitation in prior research on moral outrage, no studies have directly compared the role of moral outrage to anger, disgust, and contempt in response second- versus third-party norm violations.

By concomitantly examining the emotional underpinnings of 2PP and 3PP in a common experimental paradigm, we sought to identify the affective forces that are associated with these two forms of punishment and provide more clarity on the cognitive and affective pressures that give rise to punishment behavior. Here we used a variation of the "Investment/Trust" game which allowed us to enact the same norm violation in both a second- and third-party context. Because non-cooperation in economic games may reflect selfish behavior rather than the violation of an implicit norm of returning a fair amount (Dunning et al., 2014; van Kleef, Wanders, Stamkou, & Homan, 2015) we implemented a specific norm violation by having the (fictitious) Trustee explicitly express a willingness to cooperate fairly, which was then reneged upon.

Using this paradigm we sought to address whether moral outrage is an emotional state that is uniquely expressed in a 3PP context or whether it is also prevalent in 2PP. Based on our recent finding that moral outrage uniquely mediates 3PP as well as evidence that anger is evoked by norm violations against oneself (Hutcherson & Gross, 2011; Molho et al., 2017) we hypothesize that expression of anger (measured as the proportion of individuals who select anger as their primary emotion) will be greater for second- versus third-party violations, while a greater proportion of individuals will select moral outrage in response to third- versus second-party violations. Because other emotions have also been implicated in 2PP and 3PP, we also assessed whether the reported expressions of disgust, contempt, and sadness differed under second- and third-party contexts.

2. Methods

We recruited 360 subjects (52% male; median age 32 years) from the United States via Amazon Mechanical Turk (MTurk; Burhmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013). Participants played one round of a widely used economic game (the "Investment Game"), in which they experienced a norm violation in either a second- or third-party context and subsequently made a punishment decision. The Vanderbilt University Institutional Review Board approved the experimental protocol and all participants provided informed consent. Our sample size was based on a power analysis for a test of equality of proportions for independent samples, which indicated that we would need approximately 130 participants per group (second-party and third-party) to detect a small to medium effect size (Cohen's $h = 0.35$).

In the standard two-party Investment Game one player, labeled the Investor, is bestowed points. Investors have the option of investing none, some, or all of these points with another player referred to as the Trustee, with all invested points being multiplied threefold. The Trustee then has the option to give any number of the tripled points back to the Investor, or they can keep as many of these points as they like, without any obligation beyond social norms to return some of the investment to the Investor (Berg, Dickhaut, & McCabe, 1995). The Investor is then told that they can opt to pay to reduce the Trustee's earnings, which serves as a measure of 2PP.

In a third-party version of the Investment game, a third player, labeled the Observer, is added whose role consists of watching a round between an Investor and Trustee and subsequently modifying the Trustee's monetary gains (at the Observer's cost). The labels Investor, Trustee, and Observer are used to distinguish the different stakeholders of the task to the participants in order to reduce demand effects arising from the use of words such as 'Punisher' or 'Dictator' (Pedersen, Kurzban, & McCullough, 2013).

Participants were randomly assigned as either Investor (2PP

context) or Observer (3PP context). Investors were led to believe that they were interacting with another participant assigned as a Trustee, while Observers were told they would be watching one round between an Investor and Trustee. In reality, participants were not interacting with or observing real participants in either context.

Subjects first read instructions and responded to three initial comprehension questions to ensure they understood the task. To increase the likelihood that subjects believed they were actually interacting with other people, the study consisted of three separate stages. Subjects were told that the time between stages allowed the experimenters to collect responses from the people they were participating with before proceeding to a subsequent stage. With this paradigm, we confirmed that most participants believed they were interacting with a real person via a post-trial questionnaire designed to determine whether they were aware of the possibility that they were not interacting with a real person (Pedersen et al., 2013). Subjects were asked what they believed the study was about and whether they thought there was more to the study than meets the eye. We excluded 81 subjects who demonstrated suspicion of the deception, as determined by a researcher blind to subjects' assigned context condition and decision outcomes. The researcher coded the responses on a 1–3 scale, where 1 indicated that the subject believed deception was involved, 2 indicated that the subject thought deception was possible, and 3 indicated that the subject was unaware of any deception prior to debriefing. Subjects scoring a 1 or a 2 were excluded. Subjects recruited via MTurk may be more likely than a laboratory population to have encountered deception in an online economic game previously, accounting for the relatively high number of exclusions.

At the outset, subjects were informed that any points they had at the end of the study would be converted into a real monetary bonus payment in addition to the base payment (approximately $1 per 10 min). Although they were not informed about the bonus-points-to-dollars conversion, it amounted to a maximum of $0.40. Study time ranged between 5 and 8 min, with two 5–10 min breaks between stages.

2.1. Stage 1

Subjects were told that they would play as either the Investor or the Trustee, and would complete one round of the game with another participant. Participants were not yet informed about the 3PP condition or 'Observer' role, as we did not want knowledge of a potential third-party to influence second-party decisions. They provided their first name and several sentences on why, if selected as Trustee, the Investor should trust them with an investment. They were then told that they would be contacted within several minutes to continue once matched with another participant.

Subjects were randomly assigned to either the 2PP or 3PP context. Those in the 2PP context played as Investors, while those in the 3PP context played as Observers. The role of Trustee was in fact never assigned to subjects in either context, but they were told it was a possibility in order to lead them to believe that the Trustee was a genuine participant who had been assigned to that role. This experimental paradigm allowed us to enact a social norm violation (unfair treatment of the Investor by a Trustee) in both a second- (Investor) and third-party (Observer) context.

2.2. Stage 2

Approximately 5–10 min after the first stage, subjects received an email to continue the study and were told their assigned role (Investor or Observer).

Investors (2PP context), were told they had been paired with another subject assigned as Trustee and were provided with this Trustee's name and short statement, which indicated a willingness to cooperate. Again, the Trustee was not a real person - their name and statement were one of two counterbalanced responses from a pilot study. Importantly, this statement of a willingness to cooperate was a critical part of the manipulation, as it constituted a specific norm violation on the part of the Trustee when they later went back on this statement. Investors were then asked how many points they wanted to invest with the Trustee out of the 20 they were given.

If selected as Observer (3PP context), the role was briefly explained and subjects were told that they were not previously informed of this position out of concern that knowledge of its existence would affect how Investors and Trustees would act (e.g., they could be more likely to cooperate if they believed that a third-party was watching). Observers were also told that they would see the interactions between a pair of participants assigned as Investor and Trustee and were presented with their names and statements. In actuality neither Investor nor Trustee existed in this context, and the names and statements were the same as those used in the 2PP context, counterbalanced across subjects.

2.3. Stage 3

After another 5–10 min break, subjects received another email directing them to the final stage.

Investors (2PP) were informed what percentage of the (tripled) points they invested had been returned to them by the Trustee. The amount was manipulated to be a random value between 10% and 30% of the Trustee's point total after the investment had been tripled. We had all subjects receive an unfair return on their investment given our interest in assessing the negative emotional responses to norm violations, which would not be evoked with cooperation by the Trustee. We varied the return on investment in order to present a range of violation strengths. Subjects were then told the monetary value of their final point total, which was fixed at $0.40 regardless of how many points they had accumulated. This allowed us to control for the bonus size across participants. They also were told the monetary value of the Trustee's final total. This dollar amount was proportionally scaled to the amount of points the Trustee ended with compared to the Investor. For example, if the Trustee ended with 50 points and the Investor with 10 (~16% return of 20-point investment), the Trustee ended with $2.00 compared to the Investor's $0.40.

Observers (3PP) were informed of the number of points the 'Investor' (recall that the Investor in the 3PP context is not a real participant) had invested and the percentage of those (tripled) points that the Trustee had returned. The Investor in this condition was set to always invest all 20 points; this was, by a large margin, the modal behavior of actual participants (in the 2PP condition) and allowed us to control for the salience of the norm violation. The amount returned by the Trustee was set to be between 10 and 30% of the tripled points. The Observer was also provided with the monetary equivalent of the ending point totals with the Investor's total bonus always fixed at $0.40 and the Trustee's bonus calculated as in the 2PP context.

On the same screen as the above information, Investors and Observers were asked to provide their primary emotional state in relation to the Trustee's decision ("Which of the following emotions best describes how you feel in response to the Trustee's decision?"). Subjects chose between anger, disgust, sadness, contempt, and moral outrage; these emotions were selected based on Ginther, Hartsough, and Marois (under review). We used a forced-choice measure for the emotions due to our own pilot data and previous literature finding that having subjects rate each emotion of interest via parallel Likert scales leads to a limited ability to dissociate between emotions (Gutierrez, Giner-Sorolla, & Vasijevic, 2012; Royzman et al., 2014). Asking subjects to select a single emotion that best describes their response has been successful in differentiating between negative emotional responses even when the ratings are highly correlated (Giner-Sorolla & Chapman, 2017; Ginther, Hartsough, & Marois, under review; Hutcherson & Gross, 2011). Subjects were then asked to provide a rating of the strength with which they experienced the selected emotion on a six-point scale from "Not at all" (as 0) to "Extreme" (as 5).

On the next screen, both Investors and Observers were informed that they could adjust the Trustee's monetary payout at a 1:4 cost (for every cent spent, the Trustee's bonus was adjusted by four cents). For Investors, the cost of adjusting the score was paid out of their $0.40 bonus. Observers were told that they were receiving a bonus of $0.40 and that the cost to adjust the score would come from that total. Subjects were told they could increase or decrease the Trustee's bonus. The word “punish” was never used in instructions in order to avoid influencing behavior.

After completing the study, participants responded to a questionnaire adapted from Pedersen et al. (2013) to determine whether they were suspicious of the deception, and also provided their age and gender. Subjects were then debriefed and paid. Screenshots of the experimental paradigm are presented in the Supplementary Materials.

3. Results

Nearly 95% of Investors chose to invest all 20 points with the Trustee. We believe this rate of cooperation is due in part to the statement made by the Trustee indicating a willingness to cooperate, as previous work has found that participants are more likely to trust the Trustee if they expect reciprocity (Fetchenhauer & Dunning, 2009), though cooperation tends to be quite high for this paradigm even in the absence of such statements (ex. De Quervain et al., 2004). Investors that invested fewer than 10 points ($n = 25$) were excluded since these subjects were unlikely to experience a similarly salient norm violation as compared to those that initially trusted the Trustee (though we note that the results presented below are consistent even when we excluded all second-party Investors who did not invest the full 20 points; see Supplementary Materials). Ultimately, 254 subjects were included in analyses after the deception and investment exclusions. Both second ($n = 123$) and third ($n = 131$) -parties punished Trustees for acting unfairly (Fig. 1), and did so in fairly similar proportions: Overall, 60% of second-parties punished unfair Trustees compared to 69% of third-parties (chi-squared test for equality of proportions $\chi(1) = 2.02$, $p = 0.16$). For those who chose to punish, we did not observe a difference between 2PP and 3PP in the amount they punished Trustees (independent-samples median test, $z = 1.80$, $p = 0.181$). Punishment behavior in both contexts followed a bimodal pattern – participants tended to punish maximally or not at all. Evidently, participants punished in a similar fashion in both 2PP and 3PP contexts.

Though 2PP and 3PP subjects punished similarly, did they express similar emotional reactions to second- and third-party norm violations, particularly in regards to anger and moral outrage? Consistent with our hypothesis, moral outrage was more frequently selected for third-party than second-party violations (chi-squared tests for equality of proportions $\chi(1) = 4.59$, $p = 0.03$) while anger was more frequently selected for 2PP than 3PP contexts ($\chi(1) = 7.05$, $p = 0.008$). These findings held even when applying a conservative Yate's continuity correction for smaller sample sizes (moral outrage: ($\chi(1) = 3.89$, $p = 0.048$); anger: ($\chi(1) = 6.00$, $p = 0.014$)). Most importantly, there was a cross-over interaction between emotion selected (anger vs moral outrage) and context (2PP vs 3PP) ($\chi(1) = 9.19$, $p = 0.002$). By contrast, there was no difference in the frequency with which subjects selected other emotions; i.e. contempt, sadness, or disgust (all p's > 0.3) (Fig. 2). The frequencies of endorsement of each emotion overall and within each context are presented in Supplementary Table S1, and are visualized in Fig. S1.

Importantly, while anger and moral outrage were differentially expressed by punishment context, they were not the predominantly selected emotions. Descriptively, sadness and disgust were the predominant responses in both the 2PP and 3PP contexts. For 2PP, contempt and anger were the next most common responses, and moral outrage was the least frequent second-party emotion. These descriptive results are borne out by a chi-squared test for equality of proportions that showed a significant difference of emotions selected ($\chi(4) = 13.51$, $p = 0.01$). Follow-up pairwise tests with a Bonferroni correction for multiple comparisons showed that this result was driven by differences in proportions between moral outrage and sadness ($p = 0.05$) while the difference between moral outrage and disgust did not reach significance ($p = 0.08$). No other proportions differed between emotions in the 2PP context (all p's > 0.57; Supplementary Table S2). For 3PP, sadness and disgust were closely followed by moral outrage and contempt while anger was the least frequent response. In

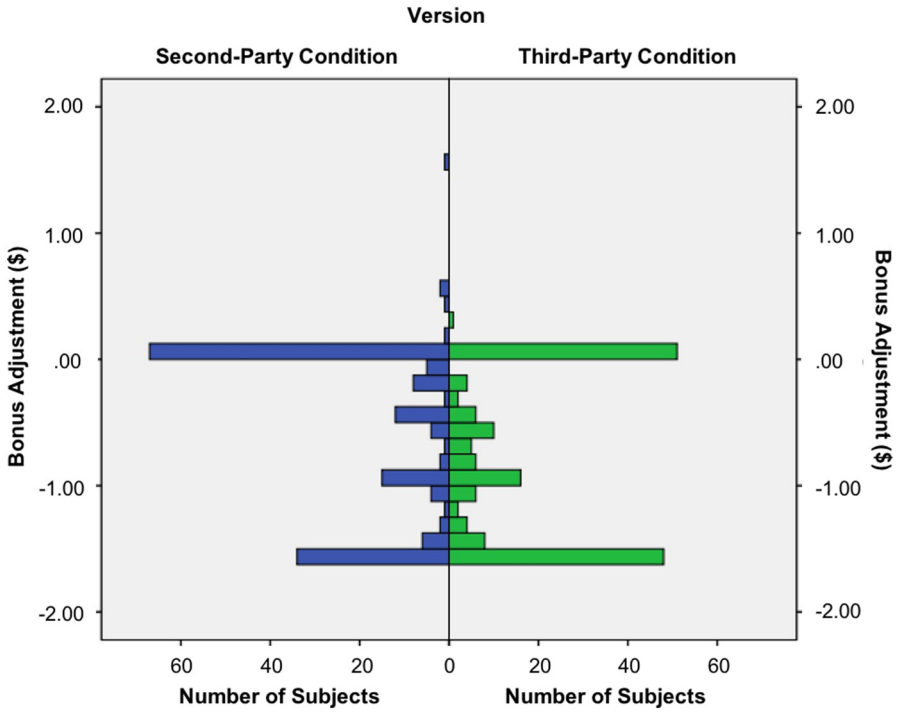

Fig. 1. Dual histograms of the amount that subjects changed the Trustee's bonus in both the 2PP and 3PP versions of the game (negative adjustment as punishment).

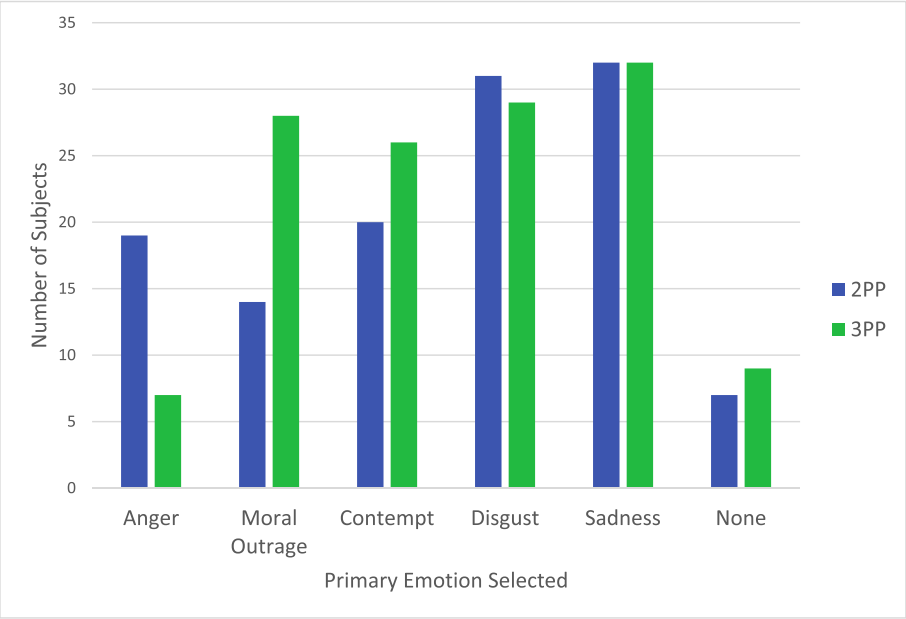

Fig. 2. Number of subjects selecting each of the emotions as their primary emotional response to the second- or third-party norm violation.

this case, the significant difference in proportions of emotions selected ($\chi(4) = 20.35, p < 0.001$) was due to anger differing from each of the other emotions (all p's < 0.008). No other proportions differed between emotions (all p's $= 1.0$; Supplementary Table S3).

We also examined subjects' mean strength ratings across emotions and contexts (2p vs 3p; Table 1). A two-way ANOVA indicated a significant main effect for selected emotion ($F(4,244) = 9.80$, $p < 0.001$), but not a significant main effect for context ($F(1,244) = 2.55$, $p = 0.11$), nor an interaction between emotion and context ($F(4,244) = 1.09$, $p = 0.36$). Bonferroni-corrected pairwise comparisons revealed that the main effect of selected emotion was due to contempt and sadness having significantly lower mean strength ratings than anger, disgust, and moral outrage (corrected p-values shown in Supplementary Table S4). Thus, there were no significant differences in the strength of emotion reported for anger and moral outrage across second and third party contexts, in contrast to their significant cross-over interaction reported for their frequency of selections (see above).

4. Experiment 2

Overall the results of this study are consistent with our initial hypothesis; there is a differential expression of anger and moral outrage – but not of other emotions – depending on punishment context even though punishment amounts and proportions remain constant: While anger is more predominantly expressed in a second-party than third-party context, the opposite is true for moral outrage.

To confirm the replicability of these results, we carried out a second experiment in a separate set of subjects taking care to pre-register the hypotheses and analyses at Open Science Framework. All methods were identical to the first experiment. We recruited a total of 375 new participants via Amazon Mechanical Turk. As in the first experiment, we excluded subjects who demonstrated a suspicion that they were not interacting with real individuals as well as those who did not invest at least 10 points with the Trustee, leaving a total of 264 participants (2P: $n = 124$, 3P: $n = 140$).

Table 1
Means and standard deviations for the strength ratings provided for each emotional response within each context.

	2p	3p
Anger	3.84 (1.12)	3.57 (0.98)
Disgust	4.00 (1.34)	3.59 (1.05)
Sadness	2.84 (1.11)	3.00 (1.29)
Contempt	3.40 (1.27)	2.81 (1.06)
Moral Outrage	3.93 (1.21)	3.61 (1.13)

We found that 67% of participants in the second-party context chose to punish compared to 62% of those in the third-party context. A chi-squared test for equality of proportions indicated that this difference was not significant ($\chi(1) = 0.66$, $p = 0.42$). For those who chose to punish, we did not observe a difference in the magnitude of the punishment across contexts (independent-samples median test, $z = 0.02$, $p = 0.98$). These findings are consistent with those of Experiment 1. Further, participants again tended to punish maximally or not at all (Fig. 3).

Critically, we replicated the cross-over relationship between anger vs moral outrage and punishment context (2P vs 3P), $\chi(1) = 5.46$, $p = 0.02$. Anger was again selected as the primary emotion at a greater proportion in the second-party versus third-party context ($\chi(1) = 5.89$, $p = 0.02$). Moral outrage was selected at a greater proportion in the third-party versus second-party context though this difference did not reach significance, $\chi(1) = 2.14$, $p = 0.14$. As in Experiment 1, disgust, contempt, and sadness were selected at similar proportions between contexts (all p's > 0.38; Fig. 4, Supplementary Table S5, Supplementary Fig. S2).

Finally, the primary emotional responses in the second-party context were sadness, anger, and disgust. A chi-squared test for equality of proportions showed significant difference of emotions selected ($\chi(4) = 16.63$, $p = 0.002$). Follow-up pairwise tests with a Bonferroni correction for multiple comparisons showed that this result was driven by difference in proportions between moral outrage and sadness ($p = 0.002$), consistent with Experiment 1 (Supplementary Table S6). For the third-party context, disgust, sadness, and contempt were the primary emotional responses. There was a significant difference in emotions selected ($\chi(4) = 13.47$, $p = 0.009$) which was due to anger being significantly lower than disgust ($p = 0.02$) again consistent with Experiment 1 (Supplementary Table S7). The emotion strength ratings again demonstrated a main effect of emotion selected ($F(4,254) = 7.30$, $p < 0.001$), and Bonferroni-corrected pairwise comparisons showed that this was due to contempt and sadness having lower mean ratings

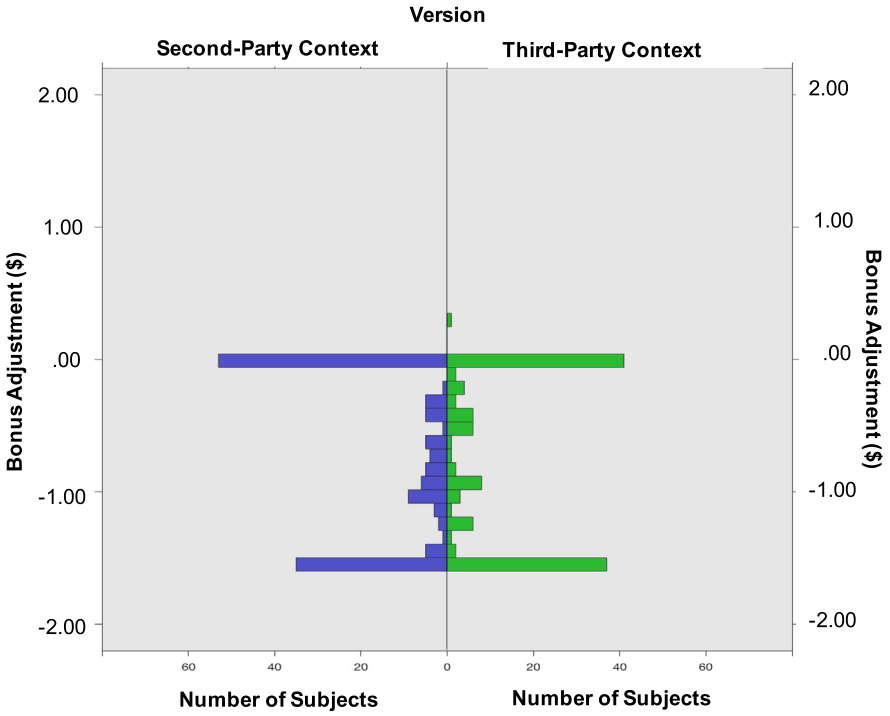

Fig. 3. Dual histograms of the amount that subjects changed the Trustee's bonus in both the 2PP and 3PP versions of the game (negative adjustment as punishment); replication study.

than the other emotions (Supplementary Tables S8 and S9).

Overall, the results of Experiment 2 provide compelling confirmatory evidence for the findings and conclusions of Experiment 1.

5. Discussion

In two experiments, we observed that a majority of subjects in both the 2PP and 3PP contexts chose to punish the Trustee's unfair behavior. Both the proportion and magnitude of punishment did not differ

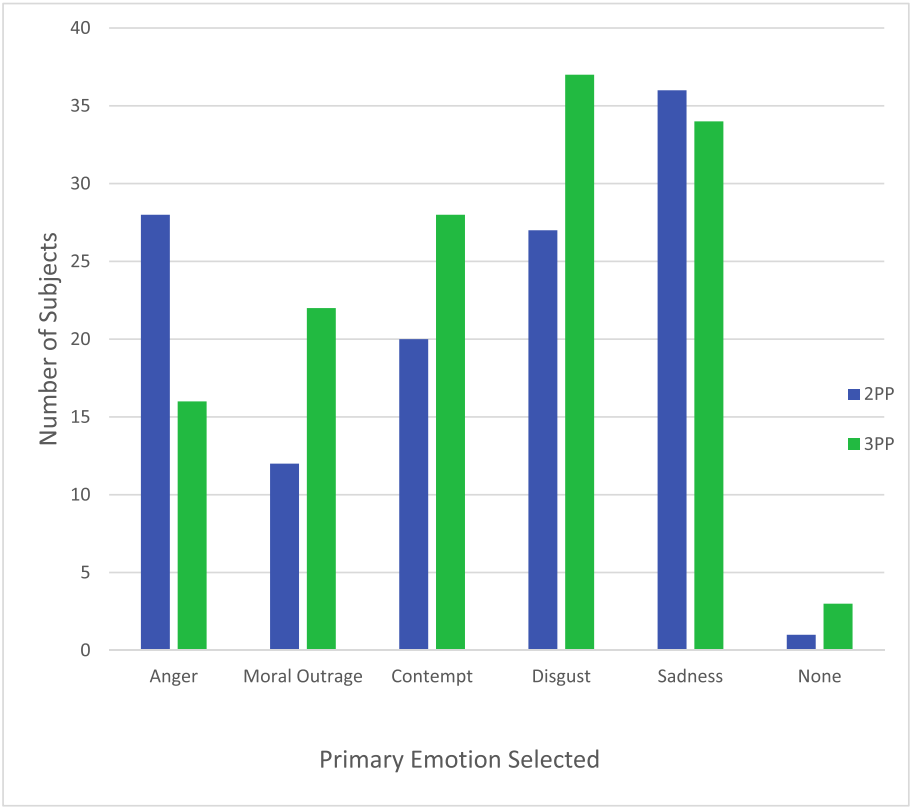

Fig. 4. Number of subjects selecting each of the emotions as their primary emotional response to the second- or third-party norm violation; replication study.

between contexts: it showed a bimodal distribution with many subjects either not punishing at all or punishing maximally. This bimodal distribution for punishment amount suggests that both second- and third-parties either refrained from punishing or sought to maximize retribution, rather than spending a token amount to signal disapproval. Our findings of comparable punishment rates for 3PP and 2PP stand in contrast to studies suggesting that 3PP is less robust than 2PP and may even simply be a product of methodological limitations (Krasnow, Delton, Cosmides, & Tooby, 2016; Kriss, Weber, & Xiao, 2016; Pedersen et al., 2013). Since the present study accounts for the methodological concerns with 3PP studies previously raised by Pedersen, Kurzban, & McCullough (2013; i.e. demand effects, audience effects, strategy method, and affective forecasting), our results support the conclusion that 3PP is robustly expressed in humans.

Most studies assessing punishment using economic games rely on the assumption that subjects expect cooperation from others (ex. Evans & Krueger, 2014; Rubinstein & Salant, 2016). However, this may not be reasonable in staged interactions with economic goals, as participants may expect others to act in their own interests (Dunning et al., 2014). In this framework, the Trustee's non-cooperation may not be perceived as a norm violation, but rather a rational economic decision reflecting normal behavior in the provided context. In contrast, our experimental design created a specific norm violation by having the (fictitious) Trustee provide a statement expressing a willingness to cooperate, which established an expectation for fairness that was then violated by the unfair return on the investment. This may account for the increased 3PP observed relative to previous economic studies of 3PP in which participants may not have perceived a norm violation, and also served to increase the emotional saliency of the unfair return. Additionally, our efforts to make subjects believe they were actually interacting with other participants helped in producing valid emotional responses to the norm violation, as emotional responses to unfairness are reduced when subjects believe they are playing against a computer rather than another person (Bone & Raihani, 2015). We acknowledge that the reliance on economic games to study complex social interactions is a limitation of this field of research, though it allows for the same norm violation to be enacted in both a second- and third-party context and informs existing psychological literature. Future work should explore different forms of norm violations in non-economic domains to determine whether these trends are consistent in regards to everyday social behavior.

Even though second- and third-party norm violations evoked similar amounts of punishment, the results of our two experiments indicate that they are associated with distinct emotional palettes. Specifically, and consistent with our prediction, we found an interaction between selected emotions and punishment context: expression of anger is more prevalent in second-party than third-party violations while expression of moral outrage is greater for third-party than second-party violations. By contrast, responses of contempt, disgust, and sadness are largely unaffected by second- and third-party contexts. These results are in line with hypotheses that moral outrage may be specific to 3P violations (Batson et al., 2007; Landmann & Hess, 2016), as well as our previous finding that moral outrage – but not anger, disgust or contempt – mediates the influence of norm violations to drive 3PP (Ginther, Hartsough & Marois, under review). By the same token, the present results also highlight the greater prevalence of anger in 2PP, consistent with previous findings that anger is evoked in response to norm violations that affect the individual and his or her goals (Batson et al., 2007; Carver & Harmon-Jones, 2009; Molho et al., 2017). Our cross-over interaction results are even more interesting when considered in the context of subjects' ratings of the strength of their expressed emotions: In both experiments, participants reported experiencing anger and moral outrage at similar strengths, regardless of the context (2PP or 3PP). This is consistent with the view that arousal is a separate dimension of affect from its valence/category (e.g. Colibazzi et al., 2010; Kesinger & Corkin, 2004), and is congruous with prior research showing high correlations between strength ratings of various negative valence emotions (e.g. Giner-Sorolla & Chapman, 2017; Hutcherson & Gross, 2011).

Importantly, our results do not support the proposition that there is a single emotion that predominates in response to a norm violation, either in the 2P or 3P context. There was no statistically significant difference in the proportion of subjects that selected contempt, disgust, and sadness as compared to moral outrage in the 3P context. Similarly, a large proportion of subjects chose emotions other than anger in the 2P context, and there was no significant difference in the proportion of subjects that selected anger, contempt, disgust, or sadness in this context. Sadness was the predominant response in both the 2P and 3P contexts (though not significantly so). Sadness may imply a focus on the loss of potential earnings, rather than being directed at the Trustee. It appears that a high proportion of subjects responded primarily to the loss itself regardless of the context. Disgust and contempt are thought to reflect character judgments (Hutcherson & Gross, 2011); the norm violation in the current study involved not only unfair treatment (low ROI) but also the initial statement by the Trustee to cooperate. Subjects may be responding to the deception by the Trustee with endorsement of disgust (and contempt to a lesser degree) signaling a character judgment.

The variety of primary emotions selected by participants points to the importance of individual differences in emotional responses to prohibited acts, with these individual differences perhaps revealing distinct internal motivations. For example, anger has been found to lead to punishment aimed at changing another's behavior directly, whereas disgust contributes to punishment in the form of avoidance and ostracism (Carver & Harmon-Jones, 2009; Molho et al., 2017). Trait-level individual differences may in turn underpin different emotional responses to norm violations. Justice sensitivity is one such trait that can predict how individuals experience situations as unjust, as well as the intensity of their emotional responses to norm violations, and their willingness to act in order to restore justice (Schmitt, Neumann, & Montada, 1995). Individual differences in observer justice sensitivity are associated with following social norms and concerns about justice in general; individuals with high levels of observer justice sensitivity tend to engage in more altruistic behavior, including altruistic (3P) punishment, and observer justice is associated with moral outrage (Lotz et al., 2011). Victim justice sensitivity is more focused on self-relevant justice concerns and is associated with greater emotional responses to perceived injustice against oneself, particularly anger (Fetchenhauer & Huang, 2004; Lotz, Okimoto, Schlosser, & Fetchenhauer, 2011). Future research could examine how these trait differences may predict punishers versus non-punishers in both second- and third-party contexts. For instance, those who are high in victim justice sensitivity may be more likely to engage in punishment in a second-party context due to experiencing greater anger, but may not engage in third-party punishment that is costly to them. Assessing and comparing the cognitive motivations and behavioral outcomes for moral outrage, anger, disgust, contempt, and sadness could contribute to distinguishing between these emotional states and their role in different forms of punishment and their relation to different forms of justice sensitivity. Finally, it is also worth mentioning that while our use of the forced-choice method has the advantage of differentiating between emotions, it limits our ability to assess whether individuals experienced multiple emotions simultaneously in response to the norm violation. It will be important for future experiments to address this limitation.

Though our results do not point to a unique emotional state that can be associated with 2P or 3P norm violations, they do provide important insights into what differentiates the expression of moral outrage and anger in the context of norm violations. Specifically, our findings provide support for the interpretation that moral outrage reflects a distinct emotional response from anger evoked by third-party norm violations. The adjective “moral” increases the perceived relevance of moral transgressions (Russell, Piazza, & Giner-Sorolla, 2013)- it is possible that this contributes to the endorsement of the emotional construct of

moral outrage in 3P contexts which may be viewed in a moral framework in comparison to violations against oneself. Viewed in this context, it is possible that the word "moral" inflated the endorsement of moral outrage in the 3P context. It would be valuable in future work to manipulate the association of the adjective 'moral' to all of the emotions across both 2P and 3P contexts to determine whether it would differentially affect emotion selection solely in the third-party context. Darley and Pittman (2003) suggested that moral outrage is a product of both the cognitive interpretation of, and the emotional response to an event. It may be that this cognitive interpretation involves interpreting actions within a moral context, which leads to moral outrage being expressed largely in response to third-party norm violations. Future work could ask participants whether they believe a behavior is morally right or wrong in addition to collecting emotional responses to further conceptualize the role of moral outrage in punishment behavior. We believe our findings support the idea that moral outrage is a distinct emotional response from anger, as many theories of emotion differentiate emotional states based on the specific circumstances that evoke the response and the behaviors associated with the emotion (Cameron, Lindquist, & Gray, 2015; Hutcherson & Gross, 2011; Moors, Ellsworth, Scherer, & Frijda, 2013). Were moral outrage and anger the same affective response, we would have expected subjects to endorse them equally, even within the same punishment context. Importantly, here we demonstrate that anger and moral outrage are evoked at different rates by different contexts. Finally, previous work from our lab showed that moral outrage mediates 3PP via the interaction of culpability and harm severity, while anger does not (Ginther, Hartsough, & Marois, under review). Taken together, these findings suggest that anger and moral outrage, while similar in many aspects, are evoked under distinct circumstances and can lead to different behavioral outcomes. Anger may be evoked specifically in response to perceived violations against the self, while moral outrage uniquely responds to violations committed against others (Batson et al., 2007; Lotz et al., 2011). Such specific association of an emotional state to the violation of social norms may have been a key driving force in the development of the uniquely human willingness to engage in the costly punishment of third-parties that undergirded the propagation of large-scale societal cooperation.

Author contributions

MG and RM developed the study concept and design. MG and LH collected the data. LH analyzed the data under supervision of RM and MG. All authors contributed to the writing of the manuscript.

Funding

This work was supported by a Vanderbilt University Discovery grant.

Data availability statement

The raw data supporting the conclusions of this manuscript are available at https://osf.io/8tm3b.

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2020.103060.

References

Batson, C. D., Kennedy, C. L., & Nord, L. A. (2007). Anger at unfairness: Is it moral outrage? *European Journal of Social Psychology, 37*(6), 1272–1285. https://doi.org/10.1002/ejsp.434.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122–142.

Bone, J. E., & Raihani, N. J. (2015). Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior, 36*(4), 323–330.

Bowles, S., & Gintis, H. (2011). A cooperative species: Human reciprocity and its evolution. *Journal of Economics, 104*, 191–194.

Burhmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science,* 4.

Cameron, C. D., Lindquist, K. A., & Gray, K. (2015). A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review, 19*(4), 371–394.

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

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: Evidence and implications. *Psychological Bulletin, 135*, 183–204.

Colibazzi, T., Posner, J., Wang, Z., Gorman, D., Gerber, A., Yu, S., & Russell, J. A. (2010). Neural systems subserving valence and arousal during the experience of induced emotion. *Emotion, 10*, 377–389.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3).

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

De Quervain, D., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science, 305*, 1254–1258.

Dunning, D., Anderson, J. E., Schlosser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology, 107*(1), 122–141.

Eriksson, K., Strimling, P., Andersson, P. A., & Lindholm, T. (2016). Costly punishment in the ultimatum game evokes moral concern, in particular when framed as payoff reduction. *Journal of Experimental Social Psychology, 69*, 59–64.

Evans, A. M., & Krueger, J. I. (2014). Outcomes and expectations in dilemmas of trust. *Judgment and Decision making, 9*(2), 90–103.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*(4), 185–190.

Fehr, E., & Gachter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 136–140.

Fetchenhauer, D., & Dunning, D. (2009). Do people trust too much or too little? *Journal of Economic Psychology, 30*, 263–276.

Fetchenhauer, D., & Huang, X. (2004). Justice sensitivity and distributive decisions in experimental games. *Personality and Individual Differences, 36*(5), 1015–1029. https://doi.org/10.1016/S0191-8869(03)00197-1.

Giner-Sorolla, R., & Chapman, H. A. (2017). Beyond purity: Moral disgust toward bad character. *Psychological Science, 28*(1), 80–91.

Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition, 133*, 97–103.

Gutierrez, R., Giner-Sorolla, R., & Vasijevic, M. (2012). Just an anger synonym? Moral context influences predictors of disgust word use. *Cognition and Emotion, 26*, 53–64.

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin, 44*(12), 1697.

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social-functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology, 100*(4), 719–737.

Kesinger, E. A., & Corkin, S. (2004). Two routes to emotional memory: Distinct neural processes for valence and arousal. *PNAS, 101*(9), 3310–3315.

van Kleef, G. A., Wanders, F., Stamkou, E., & Homan, A. C. (2015). The social dynamics of breaking the rules: Antecedents and consequences of norm-violating behavior. *Current Opinion in Psychology, 6*, 25–31.

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.

Kriss, P. H., Weber, R. A., & Xiao, E. (2016). Turning a blind eye, but not the other cheek: On the robustness of costly punishment. *Journal of Economic Behavior and Organization, 128*, 159–177.

Landmann, H., & Hess, U. (2016). What elicits third-party anger? The effects of moral violation and others' outcome on anger and compassion. *Cognition & Emotion,* 1–15. https://doi.org/10.1080/02699931.2016.1194258.

Lotz, S., Okimoto, T. G., Schlosser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of Experimental Social Psychology, 47*, 477–480.

Molho, C., Tybur, J. M., Guler, E., Balliet, D., & Hofmann, W. (2017). Disgust and anger relate to different aggressive responses to moral violations. *Psychological Science,* 1–11.

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review, 5*(2), 119–124.

Nabi, R. L. (2002). The theoretical versus the lay meaning of disgust: Implications for emotion research. *Cognition & Emotion, 16*(5), 695–703. https://doi.org/10.1080/02699930143000437.

Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society B: Biological Sciences, 280*(1758).

Plutchik, R. (1980). The nature of emotions. *American Scientist, 89*, 344–350.
Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *PNAS, 109*(37), 14824–14829.
Royzman, E., et al. (2014). CAD or MAD? Anger (not disgust) as the predominant response to pathogen-free violations of the divinity code. *Emotion, 14*(5), 892–907. https://doi.org/10.1037/a0036829.
Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology, 76*(4), 574–586.
Rubinstein, A., & Salant, Y. (2016). "Isn't everyone like me?": On the presence of self-similarity in strategic interactions. *Judgement and Decision Making, 11*(2), 168–173.
Russell, P. S., & Giner-Sorolla, R. (2013). Bodily moral disgust: What it is, how it is different from anger, and why it is an unreasoned emotion. *Psychological Bulletin, 139*(2), 328–351.
Russell, P. S., Piazza, J., & Giner-Sorolla, R. (2013). CAD revisited: Effects of the word moral on the moral relevance of disgust (and other emotions). *Social Psychological and Personality Science, 4*(1), 62–68.
Salerno, J. M., & Peter-Hagene, L. C. (2013). The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science, 24*(10), 2069–2078. https://doi.org/10.1177/0956797613486988.
Schmitt, M., Neumann, R., & Montada, L. (1995). Dispositional sensitivity to befallen injustice. *Social Justice Research, 8*(4), 385–407. https://doi.org/10.1007/BF02334713.
Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). *The "Big Three" of morality (autonomy, community, divinity) and the "Big Three" explanations of suffering.* Psychology Press: Morality and Health119–169.
Small, D. A., & Loewenstein, G. (2005). The devil you know: The effects of identifiability on punitiveness. *Journal of Behavioral Decision Making, 18*, 311–318.
Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: Neural and genetic bases of altruistic punishment. *NeuroImage, 54*, 671–680.
Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. L., Ginther, M. R., et al. (2014). Corticolimbic gating of emotion-driven punishment. *Nature Publishing Group, 17*(9), 1270–1275. https://doi.org/10.1038/nn.3781.
Van de Vyver, J., & Abrams, D. (2015). Testing the prosocial effectiveness of the prototypical moral emotions: Elevation increases benevolent behaviors and outrage increases justice behaviors. *Journal of Experimental Social Psychology, 58*, 23–33.