Emotion

Moral Outrage Drives the Interaction of Harm and Culpable Intent in Third-Party Punishment Decisions

Matthew R. Ginther, Lauren E. S. Hartsough, and René Marois

Online First Publication, March 4, 2021. http://dx.doi.org/10.1037/emo0000950

CITATION

Ginther, M. R., Hartsough, L. E. S., & Marois, R. (2021, March 4). Moral Outrage Drives the Interaction of Harm and Culpable Intent in Third-Party Punishment Decisions. *Emotion*. Advance online publication. http://dx.doi.org/10.1037/emo0000950

AMERICAN
PSYCHOLOGICAL
ASSOCIATION

Emotion

© 2021 American Psychological Association
ISSN: 1528-3542

https://doi.org/10.1037/emo0000950

Moral Outrage Drives the Interaction of Harm and Culpable Intent in Third-Party Punishment Decisions

Matthew R. Ginther, Lauren E. S. Hartsough, and René Marois
Department of Psychology, Vanderbilt University

The willingness of humans to engage in third-party punishment (TPP)—a lynchpin of our society—critically depends on the interaction between the wrongdoer's intent and the harm that he caused. But what compels us to punish such individuals when we are unaffected by their harms? Inconsistent with the idealized notion that TPP decisions are based on purely cognitive reasoning, intended harmful acts elicit strong emotional reactions in third-party decision makers. While these emotional responses are now believed to be a driving force in TPP decision making, there is debate about what emotions may be motivating this behavior. Here we show that—unlike anger, contempt, and disgust—moral outrage is evoked by the integration of culpable intent and severe harm, and that the expression of moral outrage alone mediates the relationship between this integrative process and punishment decisions. Sadness had the opposite effect of dampening punishment in response to accidental harms. We take these findings to indicate that moral outrage expresses the interaction of intent and harm in driving third-party punishment behavior.

Keywords: third-party punishment, emotion, moral outrage, harm, mental state

Supplemental materials: https://doi.org/10.1037/emo0000950.supp

Punishment of social norm violations is essential to human cooperation. Third-party punishment (TPP)—that is, punishment administered by a neutral party—provides both specific and general deterrence against actual and potential norm defectors, respectively. This punishment, or threat thereof, serves as a groundwork for our species' hypersociality among both kin and nonkin, and thus is thought to be a major factor underlying our unparalleled social, technological, and economic achievement (Buckholtz & Marois, 2012; Fehr & Fischbacher, 2004). A prominent instantiation of TPP today is evidenced in our modern criminal justice system, which has institutionalized this behavior into a defining element of society. It follows that studying this phenomenon is important not only because it is unique to human behavior and ubiquitous to all cultures, but also because we allocate vast sums of resources and opportunity costs toward this very behavior in the form of our modern justice system (Kyckelhahn, 2015).

While the adaptive benefits of TPP have been well-theorized and studied, much remains to be learned about the proximate factors that drive TPP behavior. This is a particularly salient question considering that TPP is often carried out in the absence of concrete and immediate benefits to the punisher and oftentimes even at a cost (Fehr & Gächter, 2002). Understanding the proximal causes of TPP requires not only identifying the external factors that trigger punishment behavior, but also elucidating the source of internal motivation to neutral parties' decision to punish in response to these external factors. By now, much is understood about the external elements of norm violations that evoke punishment: namely, the severity of the harm caused and the extent to which that harm was produced with intent (Carlsmith et al., 2002; Cushman, 2008), consistent with real-world legal practice and doctrine (LaFave et al., 1986; Shen et al., 2011). Specifically, it is particularly the superadditive interaction between culpable intent and substantial harm that leads to punishment (Cushman, 2008; Treadway et al., 2014). After all, we do not typically punish, nor feel the need to punish, bad deeds if they occur purely accidentally nor the desire to harm another if no action is taken to do so.

But confronted with a reprehensible act committed willfully, what drives a third-party observer to respond with punishment in the absence of immediate personal benefit? While we may all aspire to implement TPP through cold-headed reasoning—this ideal is often a key rationale for the establishment of an uninvolved and impartial adjudicator in our justice system—there is much evidence to suggest that it is strongly subject to emotional influence (Darley et al., 2000; Gummerum et al., 2016; Salerno & Peter-Hagene, 2013). Indeed, it has been suggested that differences in emotional arousal in response to a crime can strongly predict the administered punishment (Buckholtz et al., 2008), in line with the widely held notion that emotions are powerful drivers of adaptive behaviors such as TPP (Plutchik, 1980).

Lauren E. S. Hartsough https://orcid.org/0000-0002-5182-5083

Matthew R. Ginther is currently a judicial law clerk at the U.S. District Court for the District of Massachusetts, Boston, Massachusetts, United States.

Matthew R. Ginther and Lauren E. S. Hartsough contributed equally to this work.

Correspondence concerning this article should be addressed to Lauren E. S. Hartsough, Department of Psychology, Vanderbilt University, PMB 407817, 2301 Vanderbilt Place, Nashville, TN 37240, United States. Email: lauren.hartsough@vanderbilt.edu or laharts12@gmail.com

While it is well established that affective responses to intended harms play a key role in punishment behavior, it is less clear which emotions may actually drive that behavior. Early work proposed the "CAD" (Contempt, Anger, and Disgust) triad hypothesis, which describes emotional responses as being paired with specific social norm violations (Shweder et al., 1997). This hypothesis suggests that contempt is elicited specifically in response to violations of community standards (i.e., hierarchy), anger by autonomy (i.e., individuals' rights) violations, and disgust by violations of divinity (i.e., purity). While this model was supported by subject responses in one study (Rozin et al., 1999), the hypothesized CAD associations have otherwise been inconsistent (Hutcherson & Gross, 2011; Russell et al., 2013). Anger and disgust respond similarly to violations of autonomy and divinity, and it has been suggested that disgust may be synonymous to anger in response to moral violations (Nabi, 2002; Royzman et al., 2014). Similarly, expression of both anger and disgust strongly predict punishment severity, and their relative contributions to punishment behavior are often difficult to distinguish due to their similarity (Gutierrez et al., 2012; Piazza et al., 2013). That said, it has been suggested that anger tends to focus on the circumstances surrounding a norm violation and is expressed in order to change the target's behavior, while disgust and contempt focus more on judging character and are expressed to establish the target's reputation (Giner-Sorolla & Chapman, 2017; Hutcherson & Gross, 2011). Indeed, anger is more flexible in response to mitigating circumstances than disgust, especially in regards to the intent of the individual (Landmann & Hess, 2017; Russell & Giner-Sorolla, 2011). Disgust is also more likely to focus on judgments involving descriptions of bodily harm (Russell & Giner-Sorolla, 2011). Sadness is also often expressed in response to learning about harm to others. However, expression of sadness has not been found to predict punishment though it may contribute to a preference for victim compensation and feelings of empathy (Adams & Mullen, 2015; Lotz et al., 2011; Skorinko et al., 2014). Further, sadness may blunt the effect of other emotions such as anger on punishment decisions (Winterich et al., 2010).

The emotional experience of moral outrage has also emerged as a potential key player in punishment behavior. It is typically characterized as negative affect directed toward another in response to a norm violation, and has been found to strongly predict punishment decisions (Bastian et al., 2013; Carlsmith et al., 2002; Lotz et al., 2011; Salerno & Peter-Hagene, 2013) and to mediate the effects of offense severity and mitigating circumstances (i.e., facts that lessened the actor's culpability) on punishment decisions (Carlsmith et al., 2002). How moral outrage differs from other emotions, however, remains poorly defined. While moral outrage has been proposed to represent the combined experience of disgust and anger (Salerno & Peter-Hagene, 2013; Tetlock et al., 2000), it is unlikely to be reduced to the simple product of these two emotions. Moral outrage has been hypothesized to be distinct from anger because it is specifically evoked in response to third-party norm violations while anger is evoked in response to second-party violations (Batson et al., 2007; Landmann & Hess, 2017). Correspondingly, we recently found that anger was expressed more frequently in response to second- versus third-party norm violations, while moral outrage was expressed more frequently for third- versus second-party violations (Hartsough et al., 2020). Moral outrage has also been proposed to differ from disgust and contempt because it motivates punishment and other "direct approach" responses to change behaviors, whereas disgust and contempt motivate shunning and "indirect avoidance" responses (Carver & Harmon-Jones, 2009; Molho et al., 2017; Van de Vyver & Abrams, 2015). Consistent with theories arguing that emotions can be differentiated from one another if they are evoked by distinct stimuli and/or lead to different behavioral responses (ex. Barrett, 2006; Cameron et al., 2015; Moors et al., 2013), the specific contexts in which moral outrage is expressed define its distinctiveness. As a whole, then, these studies point to moral outrage as being perhaps best conceptualized as a complex negative affect response that serves to reinforce prosocial behavior. But because studies investigating moral outrage have not experimentally assessed it in comparison to other emotions, the distinctive importance of moral outrage in punishment decision making relative to contempt, anger, and disgust remains unknown. As mentioned above, Salerno and Peter-Hagene (2013) suggested that moral outrage may be related to the combined experience of anger and disgust. However, that study did not compare the expression and influence of moral outrage to other emotions and, critically, the measure of moral outrage was admittedly conflated with desire to punish, so it remains unclear how norm violations, emotional states, and punishment decisions are interrelated.

The present study is designed to experimentally test the relationship between norm violations, TPP, and emotions. Specifically, here we aim to untangle the expression of moral outrage and other emotional constructs in response to norm violations, and to decipher the role each play in TPP. We presented subjects with textual vignettes that described scenarios containing norm violations that varied in both mental state culpability and harm severity, and recorded the subjects' emotional and punishment responses to each of these scenarios. With this data, we assessed how contempt, anger, disgust, sadness, and moral outrage each mapped onto the components of the norm violation (i.e., intent, harm, and their interaction), and measured how these emotions differentially mediate the relationship between the norm violation and the punishment decision to gauge the relative influence of each emotion in driving TPP. We expected that the expression of anger, disgust, and moral outrage would all be associated with increased TPP. We hypothesized that moral outrage would be driven by the interaction of harm and intent and would mediate the effect of this interaction on punishment based on the model proposed by Carlsmith et al. (2002). Further, we hypothesized that anger would be predicted by the intent factor, while disgust and sadness would be predicted primarily by the severity of the harm.

Method

Participants

All subjects provided informed consent and the experimental protocol was approved by the Vanderbilt University Institutional Review Board. We recruited a total of 455 participants via Amazon Mechanical Turk. Of these, 387 (age 19 to 76 years, $M = 37$, $SD = 12$; 52% male) were included in the analysis as the others failed to complete the full survey or incorrectly answered an attention check question (see below). Sample size was chosen based on a power analysis indicating that roughly 400 participants were needed to obtain a power of 0.95 with a moderate effect size (.25) for main effects and the interaction (a moderate effect size is consistent with prior studies examining mediation effects of emotion on punishment). We initially recruited more than 400 participants (i.e., 455) because of the expectation, based on our prior studies, that a number of them would fail the attention check. Subjects were recruited from across the United States, and the use of Amazon Mechanical Turk allowed us to obtain a larger and more diverse sample of the U.S. population across a number of demographic factors than we would have had we recruited undergraduate students (Stewart et al., 2015). Most participants completed the survey in 5–8 min, for which they were compensated $0.40.

Design

Participants were randomly assigned to respond to one test scenario depicting the actions of a protagonist named "John" that resulted in harm to another person. Prior to the test scenario, they completed three unanalyzed practice scenarios that introduced the task design and presented the full spectrum of possible harm and mental state levels.

The experiment employed a 4 (Harm Level) × 4 (Mental State Level) between-subjects design. Each participant read a single test scenario that included one of four possible harm severities (negligible, moderate, severe, or death) caused by John's actions and one of four possible mental states (purposeful, reckless, negligent, or blameless) John could be in when committing the act. These mental states are utilized and defined in the Model Penal Code's mental state hierarchy; briefly stated, purposeful indicates that the individual acted with the intention of causing the harm, reckless means that the individual committed the act despite being aware of the substantial risk that the harm would occur, negligence occurs when the individual should have been aware of the substantial risk that their actions would lead to harm, and blameless corresponds with accidental harms that are outside of the individual's control.

The test scenario was derived from one of 64 different scenario stems previously used by Ginther et al. (2014, 2016), with each stem describing a specific set of events. Each of the 64 stems corresponded to one of the four levels of harm caused by John, with 16 scenario stems for each of the four harm levels. Each of the individual 64 stems could vary among the four different levels of John's mental state in causing the harm. With 64 different stems, each having four different possible mental states, there was a total of 256 scenarios from which the test scenario was randomly sampled for each subject. We presented a single scenario to each subject so as to ensure that the background information in any given scenario did not unduly influence decisions. Our large sample size allowed us to present a range of scenarios across participants, negating the influence of scenario context on punishment decisions.

Each scenario was presented in three phases. The first phase was an introduction sentence that provided relevant background information about the scenario. The second and third phases presented John's mental state and the resulting harm, with the order of presentation of the mental state and harm information counterbalanced across subjects (i.e., mental state in the second phase was followed by harm in the third phase, or vice versa). These hypothetical scenarios allowed us to portray a range of realistic situations while carefully manipulating mental state and harm across a range of levels. Subjects were not told that the scenarios depicted hypothetical situations only. Tables S1a and S1b in the online supplementary materials provide sample scenarios. Subjects' progression through these steps was self-paced.

Emotional Responses

Subjects' emotions were assessed as described below at two points during each scenario: after the presentation of the second phase (i.e., after presentation of either harm or mental state information) and again after the presentation of the third phase (i.e., after presentation of both harm and mental state information). The purpose of the first emotional assessment was to isolate the emotional responses to harm and mental state independent of one another, whereas the second assessment was used determine the emotional responses to the integration of harm and mental state.

Previous findings suggest that parallel Likert scales (i.e., having subjects provide a rating for each emotion of interest) are limited in their ability to dissociate between measures, including emotions (Gutierrez et al., 2012; Royzman et al., 2014). In particular, studies seeking to directly compare anger and disgust have found these two emotions to be largely indistinguishable in response to social norm violations, particularly due to their ratings being highly correlated when those are collected using parallel Likert scales (Gutierrez et al., 2012; Nabi, 2002). A means to circumvent this problem and differentiate between emotions is to ask subjects to select the single emotion that best describes their response to a given stimulus; this approach has been successful even when the ratings provided for each emotion are highly correlated (Giner-Sorolla & Chapman, 2017; Hutcherson & Gross, 2011; Russell et al., 2013). Thus, for the current study we had subjects select the emotion (either anger, disgust, contempt, moral outrage, or sadness) that best identified their primary and secondary emotional states.

The purpose of the secondary emotional response selection was to assess whether there were specific associations between emotion selections. The secondary emotional response data, however, only revealed that anger is the predominant secondary response for all other primary emotions (see Section 5 and Figure S6 in the online supplementary materials), and the mediation pathways using the secondary emotional responses did not yield significant mediating effects or patterns (see Section 6 in the online supplementary materials). The secondary response data is therefore not discussed further to preserve focus on the study's primary aims.

After selecting each of their emotional responses, subjects were asked to rate how strongly they experienced that emotion on a 10-point Likert scale with 0 as *Not at all* and 9 as *Extreme*. The results of these ratings confirmed previous findings of high correlation between emotions when using a Likert scale (Gutierrez et al., 2012; Nabi, 2002) and are not discussed further in this paper.

Punishment Response

After subjects reported their final emotional response in the third phase of the scenario, they provided a punishment rating. They were asked to indicate how much they felt John should be punished for his behavior on a 10-point Likert scale with 0 as *No punishment* and 9 as *Most severe punishment*. This provided a measure of subjects' intuition about the appropriate amount of punishment.

Attention Check

After completing the test scenario, participants were presented with an "attention check" scenario. This scenario appeared identical to the test and practice scenarios in structure but had a sentence embedded within that told subjects to select specific responses regardless of their own response to the scenario. This allowed us to screen out individuals who did not carefully read the scenarios. Following the attention check, basic demographic information was collected and individuals were debriefed and provided with instructions for compensation.

Analyses

Our statistical analyses focused on answering two primary questions: First, what are the emotional responses to varying levels of harm and mental state, both independently and when these factors are integrated? Second, can the experience of specific emotions be linked to punishment behavior?

To address the first question, we relied on regression analyses to examine the relationships between the norm violation and emotion. For all regression analyses, predictors were standardized through *z* transformation so as to enable meaningful comparisons. When examining the effect of the norm violation on the selected emotion, we used linear probability models with each emotion expressed as a binary variable (selected or not) regressed, in independent analyses, on harm level, mental-state level, and their interaction, with heteroskedastic robust standard errors to allow for non-normality of variances around the probability estimates (Aldrich & Nelson, 1984). Lines of best fit for the predicted probability of selecting each emotion as a function of harm and mental state were generated to further visualize these relationships.

To address the second question, separate mediation analyses were run to test whether the expression of each emotion mediated the effect of harm, mental state, and the interaction of mental state and harm on subjects' punishment ratings. Emotions were again treated as binary variables (selected or not). We used a counterfactually defined causal mediation method, as the product-of-coefficients approach to calculating indirect effects is not robust for binary mediators (Imai et al., 2010; Pearl, 2012; Steen et al., 2014; Valeri & Vanderweele, 2013). We obtained estimates for the natural indirect effect for each emotion as a mediator of the effect of the harm, mental state, and their interaction on punishment using the R package Medflex (Steen et al., 2014). Standard errors were calculated using the bootstrap method with 1,000 draws (Preacher & Hayes, 2008) and 95% confidence intervals were generated for each indirect effect.

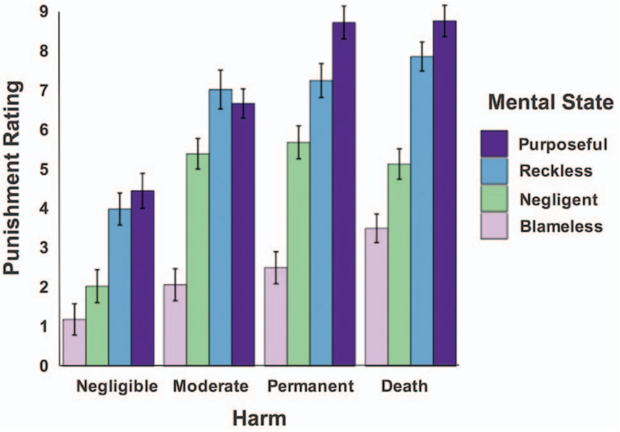
Results

Punishment

As expected, punishment behavior was characterized by not only an effect of harm (b = 0.38, se = 0.04, $p < .001$, 95% CI [0.31, 0.45]) and mental state (b = 0.60, se = 0.04, $p < .001$, 95% CI [0.53, 0.67]) but also a superadditive interaction between the two (b = 0.10, se = 0.04, $p = .01$, 95% CI [0.03, 0.17]; Figure 1), consistent with prior findings (Ginther et al., 2016). Increasing harm severity predicted increased punishment, and even more so for more culpable levels of mental state. These relationships reflect Path C in subsequent mediation analyses (see Figure 2).

Figure 1

Mean Punishment Ratings as a Function of Mental State and Harm Level



Note. Error bars display +/− 1 standard error of the mean.

Independent Effects of Harm and Mental State on Emotion Selection

How do presentations of harm and mental state—independent of one another—affect the emotional response of the participants and, ultimately, their punishment decision? To assess the first part of this question, we examined subjects' first emotion response, which occurred after the subject had been presented with information as to either the level of harm or mental state, but not both. Specifically, we tested for linear trends in subjects' emotional responses as a function of level of mental state and level of harm, independently (Path A in Figure 2).

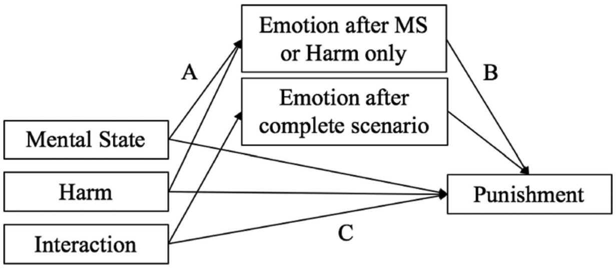
This revealed that expressions of sadness and moral outrage increased with increasing levels of harm severity while disgust and contempt decreased (note beta values in Path A column in Table 1 and see Figures S1 and S2 in the online supplementary materials for a graphical depiction). For mental state, expressions of anger, disgust, and moral outrage increased as the level of intent increased while sadness decreased (note beta values in Path A column in Table 2 and see Figures S3 and S4 of the online supplementary materials for a graphical depiction).

Using mediation analyses, we assessed which of these emotional responses to harm or mental state might demonstrate an indirect effect on punishment decisions. The mediation analyses examine the size and directionality of the indirect effects (Path A and Path B in Figure 2) in the presence of the direct effect (Path C in Figure 2). Moral outrage manifested an indirect effect for both mental state and harm, while sadness manifested an indirect effect for harm alone. No other mediation effects were significant (see Mediation Effect columns in Tables 1 and 2). The B and C Paths are included as Tables S2 (harm) and S3 (mental state) in the online supplementary materials.

Thus, while various emotions tracked increasing levels of either mental state or harm, only moral outrage increased commensu-

Figure 2

Graphical Depiction of Regression and Mediation Analyses Performed in the Present Study


Emotion after MS or Harm only
Emotion after complete scenario
A
B
Mental State
Harm
Interaction
Punishment
C

Note. Independent regression analyses were performed with level of mental state, harm, and the interaction as predictors. Linear regression was performed between three different paths: (a) path A between the stimulus (either mental state, harm, or the interaction) and the emotion response, (b) path B between the emotion response and the punishment amount, and (c) path C between the stimulus and the punishment amount. Note that when examining the effect of mental state or harm alone, the emotion selected following the presentation of that respective predictor (mental state or harm) was used. For the interaction we used the emotion selected after subjects evaluated the complete scenario. In a separate analysis, in order to determine the presence of a mediation effect, for each set of predictors (mental state, harm, and their interaction) we examined the size of the indirect effect (a * b[1]) in the presence of the direct effect (c).

rately with both increasing levels of harm and mental state. Further, only moral outrage and sadness displayed a mediating effect between the triggering factors (harm, mental state) and the punishment response. We note that while these analyses allow us to assess the emotional responses to harm and mental state independent of one another (by using the first emotional response made after only harm or mental state has been presented), the mediation effect uses the punishment decision made after both harm and mental have been presented. This somewhat limits our ability to interpret the independent mediating effect of each emotion for harm or mental state onto punishment, and our use of the second emotional response in the following analyses (after both harm and mental state were presented) is essential to investigate the emotions involved in the interaction of harm and mental state.

Interactive Effect of Harm and Mental State on Emotion Selection

In addition to examining the characteristics of mental state and harm independently, we assessed the integration of mental state and harm and the commensurate effect on punishment by repeating the above analyses on the second (final) emotional response (i.e., the response provided after the evaluation of both mental state and harm). While before we tested for linear trends as a function of mental state and harm independently, here we examine a linear trend with the interaction of mental state and harm.

Moral outrage, contempt, and sadness each demonstrated an interaction between the mental state and harm factors (see Path A column in Table 3 and Figure 3 for a graphical depiction). Importantly, we only observed a superadditive interaction—paralleling the punishment behavior—in the case of moral outrage: there was no effect on anger and disgust, and the interaction of harm and mental state negatively predicted contempt and sadness. In other words, subjects experienced sadness primarily in response to unintentional harms, and the likelihood of experiencing sadness scaled with the severity of the harm (Figure 3 and Figure S5 in the online supplementary materials). The interaction of harm and mental state also negatively predicted contempt, with the interaction effect driven by a large proportion of subjects experiencing contempt for negligent accidents that resulted in low severity harms (see Figure 3). Figure S5 in the online supplementary materials shows the proportions of each emotion selected for different levels of the interaction of harm and mental state.

The above analyses indicate that moral outrage was the only emotion sensitive to both increasing harms and culpable intent, and that moral outrage displayed a superadditive interaction effect in relation to mental state and harm that mirrors the superadditive effect driving subjects' punishment decisions (as in Figure 1). To determine whether the expression of moral outrage mediated the effect of the interaction of harm and mental state on the superadditive punishment behavior, we assessed whether moral outrage— or any other emotion— mediated the effect of the interaction of harm and mental state on subjects' punishment decisions. Mediation effects were found only for moral outrage and sadness, though they mediated the opposite effects on punishment (see Mediation Effect column Table 3; B and C paths are presented as Table S4 in the online supplementary materials). Specifically, subjects were more likely to experience moral outrage where there were high levels of harm and culpable mental states, but less likely to experience sadness. Furthermore, expression of moral outrage was associated with greater punishment while the expression of sadness was associated with lesser punishment. Finally, we compared the strength of this mediation effect with the strength of the mediation effect that moral outrage displayed for harm and mental state independently. The mediation effect for the interaction is substantially greater than for either harm or mental state ($Z = 2.86$, $p = .0043$; Paternoster et al., 1998).

Discussion

Our results indicate that the expression of moral outrage is uniquely critical to third-party punishment (TPP) in humans in two ways. First, we demonstrate that while most of the emotions tested respond to either increases in harm or increases in mental state, moral outrage is selectively expressed by the interaction of culpable intent and harmful outcomes. Second, we observe that the expression of moral outrage selectively mediates the effects of both harm and mental state on punishment decisions. We discuss these two links between moral outrage and punishment below.

As noted in the introduction, recent findings have observed that punishment behavior is characterized by a superadditive effect of culpable intent and harmful outcome (Ginther et al., 2016; Treadway et al., 2014). A primary goal of the present study was to

[1] As noted in the methods—we used a counterfactually defined causal mediation method in order to calculate the indirect effect due to concerns about the effect of a binary regressor (the emotion selection) on the product of coefficients approach (a * b). Nonetheless, the product of coefficients approach provides nearly identical results in this case and is helpful for conceptualizing the nature of the mediation effect.

Table 1

Relationship Between Harm, Emotion, and Punishment

	Path A: Harm to emotion				Mediation effect on punishment			
Emotion	*b*	*se*	*p*	95% CI	*b*	*se*	*p*	95% CI
Anger	−0.02	0.02	.485	[−0.06, 0.03]	−0.004	0.01	.525	[−0.02, 0.01]
Disgust	−0.07	0.02	<.001**	[−0.10, −0.04]	−0.01	0.01	.466	[−0.03, 0.02]
Contempt	−0.03	0.02	.035	[−0.07, −0.002]	0.01	0.01	.389	[−0.01, 0.02]
Sadness	0.06	0.03	<.001**	[0.01, 0.11]	−0.03	0.01	.024*	[−0.05, −0.004]
Moral outrage	0.06	0.02	.001**	[0.02, 0.09]	0.02	0.01	.032*	[0.003, 0.05]

Note. Path A: Harm to emotion column presents standardized regression coefficients of path A (see Figure 2) between the harm stimulus and the emotional response for each emotion at the first emotion response (when only the harm had been presented). The Mediation effect on punishment column presents the magnitude and statistical significance of the mediation effect on the punishment decision.
* $p < .05$. ** $p < .005$.

determine which emotion(s) may be associated with the interaction of these two components. While a number of studies have investigated the relationship between emotion and punishment decisions (Fehr & Gachter, 2002; Kogut, 2011; Laurent et al., 2014; Salerno & Peter-Hagene, 2013), few have empirically investigated the types of norm violations that induce the expression of specific emotions (Hutcherson & Gross, 2011; Rozin et al., 1999; Russell et al., 2013). Moreover, to our knowledge, none have examined the emotional response specific to the interaction of a culpable mental state and severe harms, the lynchpin of punishable behavior. Our finding that the expression of moral outrage is selective to, and predominates at, the intersection of culpable intent and harmful outcomes not only provides experimental support for models that put this emotion at the junction of serious offenses and the absence of mitigating circumstances (e.g., Carlsmith et al., 2002), it also distinguishes it from other emotions (i.e., contempt, anger, or disgust) that have previously been implicated in punishment behavior. By revealing that subjects overwhelmingly expressed moral outrage at this junction, our results indicate that moral outrage is uniquely reflective of the human emotional response to severe harms that are the result of culpable conduct.

This aspect of our results does conflict, in part, with a recent study by Landmann and Hess (2017) that found that anger ratings were predicted by moral violation regardless of the outcome caused. They concluded that this anger at the intent to commit a violation independent of the harm can be defined as moral outrage. While our findings do support the conclusion that anger is the predominant response in the case of culpable intent without a harmful outcome (see Figure S4 in the online supplementary materials), our results are inconsistent with their conclusion that moral outrage is selective to intent. This discrepancy in findings is likely due to the fact that Landmann and Hess did not explicitly gauge subjects' moral outrage but instead relied on inference to reinterpret expressed anger as moral outrage.

The second major finding of the present study is that expression of moral outrage selectively mediates augmented punishment decisions in the case of severe, intentional harms. Insofar as moral outrage appears to be associated with punishment decisions, this result is consistent with two prior studies (Carlsmith et al., 2002; Salerno & Peter-Hagene, 2013). However, in contrast to these two prior studies, we demonstrate that expression of moral outrage is, in this way, unique among emotions. While we found that anger and disgust predicted punishment, they did not do so by mediating the effect of intent, harm, or their interaction. Precisely how anger and disgust affect TPP independently of moral outrage remains to be determined.

Studies of punishment decision making often speak only of the emotional drivers of punishment, not of emotions that may act to suppress action (Carlsmith et al., 2002; Salerno & Peter-Hagene, 2013). Two of our results support a conclusion that sadness may operate in the latter fashion. First, we observed that sadness is the predominant response, at an 8-to-1 ratio, when a severe harm is unintentionally caused, or put another way, when an accident occurs. Second, and critically, expression of sadness induced the opposite effect to moral outrage of the norm violation on punishment; that is, sadness mediated reduced punishment ratings. Pre-

Table 2

Relationship Between Mental State, Emotion, and Punishment

	Path A: Mental state to emotion				Mediation effect on punishment			
Emotion	*b*	*se*	*p*	95% CI	*b*	*se*	*p*	95% CI
Anger	0.05	0.02	.034*	[0.004, 0.09]	0.01	0.01	.259	[−0.01, 0.02]
Disgust	0.05	0.02	.006*	[0.01, 0.08]	−0.01	0.01	.126	[−0.03, 0.004]
Contempt	−0.01	0.02	.579	[−0.04, 0.02]	0	0	.633	[−0.01, 0.01]
Sadness	−0.13	0.02	<.001**	[−0.18, −0.08]	0	0.01	.745	[−0.02, 0.02]
Moral outrage	0.04	0.02	.014*	[0.01, 0.08]	0.02	0.01	.037*	[0.001, 0.04]

Note. Path A: Mental state to emotion column presents standardized regression coefficients of path A (see Figure 2) between the mental state stimulus and each emotion at the first emotion response (when only the mental state had been presented). The Mediation effect on punishment column presents the magnitude and statistical significance of the mediation effect on the punishment decision.
* $p < .05$. ** $p < .005$.

Table 3

Relationship Between the Interaction of Mental State and Harm, Emotion, and Punishment

	Path A: Interaction to emotion				Mediation effect on punishment			
Emotion	*b*	*se*	*p*	95% CI	*b*	*se*	*p*	95% CI
Anger	−0.002	0.02	.943	[−0.05, 0.05]	0	0	.945	[−0.01, 0.01]
Disgust	−0.01	0.02	.385	[−0.05, 0.02]	−0.003	0.01	.526	[−0.02, 0.01]
Contempt	−0.05	0.01	.001**	[−0.07, −0.02]	−0.004	0.01	.529	[−0.02, 0.01]
Sadness	−0.09	0.02	<.001**	[−0.14, −0.05]	0.07	0.02	<.001**	[0.03, 0.11]
Moral outrage	0.16	0.02	<.001**	[0.11, 0.20]	0.06	0.01	<.001**	[0.03, 0.08]

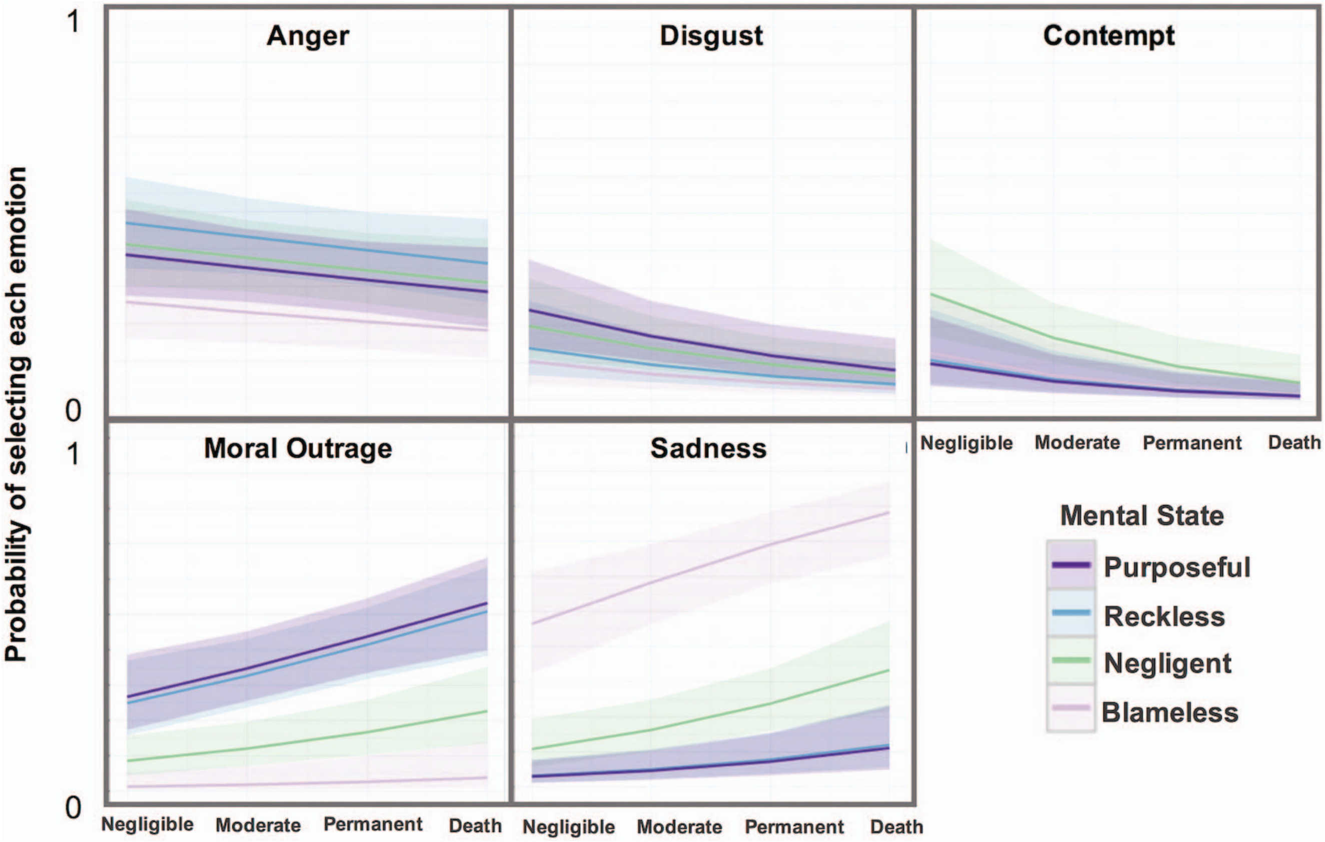
Note. Path A: Interaction to emotion column presents standardized regression coefficients of path A (see Figure 2) between the interaction term and each emotion at the second emotion response (when both mental state and harm had been presented). The Mediation effect on punishment column presents the magnitude and statistical significance of the mediation effect on the punishment decision.
** $p < .005$.

vious studies have found evidence that sadness can blunt subsequent anger and reduce the influence of anger on cognitive judgments (Winterich et al., 2010), perhaps as a result of empathy toward the offender (Skorinko et al., 2014). While further research is necessary to establish a causal link between expression of sadness and punishment suppression, this observation can potentially have substantial real-world application in both judicial and policy domains, particularly in relation to restorative justice efforts.

A key consideration in interpreting our findings is the extent to which our experimental paradigm was able to differentiate between the emotional expressions measured, as previous research has found these emotions difficult to tease apart within a TPP context (e.g., Gutierrez et al., 2012; Nabi, 2002). One potential

Figure 3

Lines of Best Fit for the Predicted Probability of Selecting Each Emotion in the Second (Final) Emotion Response as a Function of Harm and Mental State



Note. Shaded areas depict 95% confidence intervals.

methodological limitation is that we did not provide participants with definitions for the different emotional expressions they could select from, which meant we relied on their intuitive understanding of these emotions. Given our primary interest in understanding and differentiating the emotional responses of lay individuals, we felt that providing a definition for each emotion could introduce potential framing effects (e.g., associating disgust with bodily harm), especially because there remains debate even among academics over the conceptual differences in emotions (e.g., Moors et al., 2013) and selecting between definitions could have introduced bias in favor of our own hypotheses. Nonetheless, even in the absence of specific definitions, our results indicate that participants discriminated between the emotion categories, particularly for the emotions on which our main findings are based. Figure 3 and Table 3 show that our use of a forced-choice paradigm in conjunction with our parametric manipulation of harm and mental state lead to distinct emotional responses based on the context of the norm violation. Specifically, moral outrage predominated for severe culpable harm, sadness for blameless harm, and contempt tended to be selected only for negligent, low-level harms. Furthermore, anger was the predominant response for culpable mental state with low harm (see Figure S5 in the online supplementary materials). The differentiation between moral outrage and the other emotions was further borne out by the observation that moral outrage mediated the effect of the interaction of harm and mental state on punishment, while anger, disgust, and contempt did not. Previous research had not compared moral outrage and related emotions across such a range of violation contexts, and our findings suggest that the expression and differentiation of these emotions is highly context-dependent. This would be consistent with a number of theories of emotions positing that emotions are distinct if they are evoked by different stimuli and/or lead to different behavioral responses (e.g., Barrett, 2006; Cameron et al., 2015; Moors et al., 2013). Taken together with our experimental findings, these considerations support our conclusion that participants were not only able to distinguish between emotions, but also to treat moral outrage as an emotional experience distinct from the experience of anger, contempt, or disgust.

Another consideration worth pointing out is that that while our use of a forced-choice paradigm allowed us to differentiate between emotions, it limited our ability to determine whether and how emotions may be evoked simultaneously. This is especially relevant given Salerno and Peter-Hagene's (2013) conclusion that moral outrage reflects the combined experience of anger and disgust. The fact that moral outrage was the predominant response for severe intentional harms, for example, does not necessarily mean that subjects were not also experiencing other emotions, only that moral outrage was the strongest. If moral outrage is a combination of anger and disgust as suggested by Salerno and Peter-Hagene, it may be that moral outrage was selected at higher rates in response to severe culpable harms because it acted as an "emotional shortcut" to the presence of both anger and disgust. However, were this relationship that simple, we would have expected moral outrage to be selected under low culpability and low harm conditions as well, especially since this is when anger and disgust are both preferentially selected (see Figure 3). Additionally, conducting the mediation analyses with the secondary emotions (i.e., emotions experienced second-most strongly) showed that neither anger nor disgust mediated the effect of harm, mental state, or their interaction as a secondary emotion (see Section 6 in the online supplementary materials), and thus did not closely mirror the effects of moral outrage. Furthermore, we recently found that while anger was reported more frequently in response to second- versus third-party norm violations, it was the opposite for moral outrage, which was reported more frequently for third-party violations (Hartsough et al., 2020). By contrast, disgust was not endorsed at different rates between second- and third-party contexts. Taken together, these findings suggest that the distinct patterns of emotional expression we found between moral outrage, anger, and disgust are not simply a result of the limitations in reporting multiple emotions simultaneously, but rather reflect the complexity of the relationship between moral outrage, anger, and disgust. Specifically, our findings indicate that if moral outrage does represent a superordinate combination of anger and disgust, it is not simply reducible to the product of those experiences as it is contingent on the specific combination of contexts and conditions necessary to evoke it (i.e., a third-party context combined with culpable intent resulting in severe harm). That is not to say that anger or disgust may not also be expressed under such conditions, however; only that these emotions may be responding primarily to certain components of the violation (e.g., culpable intent for anger) rather than the interaction of intent and harm that leads to the expression of moral outrage.

Regardless of what the relationship between moral outrage, anger, and disgust may ultimately prove to be, the present study highlights the central role of moral outrage in driving third-party punishment. When subjects are asked to identify their emotions after evaluating a severe culpable harm, they select moral outrage over anger at a 2-to-1 ratio and disgust at a 3-to-1 ratio, suggestive of the primacy of this emotional construct. Moreover, it's the expression of moral outrage alone that mediates the relationship between culpable harms and punishment decisions. These results indicate that it is moral outrage—more so than contempt, anger, or disgust—that provides the emotional impetus for punishing third-party transgressors, befitting moral outrage's consideration as a complex negative affect toward harmful antisocial behavior.

References

Adams, G. S., & Mullen, E. (2015). Punishing the perpetrator decreases compensation for victims. *Social Psychological and Personality Science*, *6*(1), 31–38. https://doi.org/10.1177/1948550614542346

Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit, and probit models*. SAGE. https://doi.org/10.4135/9781412984744

Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, *10*(1), 20–46. https://doi.org/10.1207/s15327957pspr1001_2

Bastian, B., Denson, T. F., & Haslam, N. (2013). The roles of dehumanization and moral outrage in retributive justice. *PLoS ONE*, *8*(4), e61842. https://doi.org/10.1371/journal.pone.0061842

Batson, C. D., Kennedy, C. L., Nord, L. A., Stocks, E. L., Fleming, D. Y. A., Marzette, C. M., Lishner, D. A., Hayes, R. E., Kolchinsky, L. M., & Zerger, T. (2007). Anger at unfairness: Is it moral outrage? *European Journal of Social Psychology*, *37*(6), 1272–1285. https://doi.org/10.1002/ejsp.434

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

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655–661. https://doi.org/10.1038/nn.3087

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. https://doi.org/10.1177/1088868314566683

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., & Harmon-Jones, E. (2009). Anger is an approach-related affect: Evidence and implications. *Psychological Bulletin*, *135*(2), 183–204.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Darley, J. M., Carlsmith, K. M., & Robinson, P. H. (2000). Incapacitation and just deserts as motives for punishment. *Law and Human Behavior*, *24*(6), 659–683. https://doi.org/10.1023/A:1005552203727

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

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

Giner-Sorolla, R., & Chapman, H. A. (2017). Beyond purity: Moral disgust toward bad character. *Psychological Science*, *28*(1), 80–91. https://doi.org/10.1177/0956797616673193

Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., & Marois, R. (2016). Parsing the behavioral and brain mechanisms of third-party punishment. *The Journal of Neuroscience*, *36*(36), 9420–9434. https://doi.org/10.1523/JNEUROSCI.4499-15.2016

Ginther, M., Shen, F., Bonnie, R., Hoffman, M., Jones, O., Marois, R., & Simons, K. W. (2014). The language of mens rea. *Vanderbilt Law Review*, *67*(5), 1327–1372.

Gummerum, M., Van Dillen, L. F., Van Dijk, E., & Lopez-Perez, B. (2016). Costly third-party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *Journal of Experimental Social Psychology*, *65*, 94–104. https://doi.org/10.1016/j.jesp.2016.04.004

Gutierrez, R., Giner-Sorolla, R., & Vasiljevic, M. (2012). Just an anger synonym? Moral context influences predictors of disgust word use. *Cognition and Emotion*, *26*(1), 53–64. https://doi.org/10.1080/02699931.2011.567773

Hartsough, L. E., Ginther, M., & Marois, R. (2020). Distinct affective responses to second- and third-party norm violations. *Acta Psychologica*, *205*, 103060. https://doi.org/10.1016/j.actpsy.2020.103060

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. https://doi.org/10.1037/a0022408

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334. https://doi.org/10.1037/a0020761

Kogut, T. (2011). The role of perspective taking and emotions in punishing identified and unidentified wrongdoers. *Cognition and Emotion*, *25*(8), 1491–1499. https://doi.org/10.1080/02699931.2010.547563

Kyckelhahn, T. (2015). *Justice expenditure and employment extracts, 2012-preliminary (NCJ 248628)*. U.S. Bureau of Justice Statistics. Retrieved from https://www.bjs.gov/index.cfm?ty=pbdetail&iid=5239

LaFave, W. R., & Scott, A. W. (1986). *Criminal law*. West Publishing.

Landmann, H., & Hess, U. (2017). What elicits third-party anger? The effects of moral violation and others' outcome on anger and compassion. *Cognition and Emotion*, *31*(6), 1097–1111. https://doi.org/10.1080/02699931.2016.1194258

Laurent, S. M., Clark, B. A. M., Walker, S., & Wiseman, K. D. (2014). Punishing hypocrisy: The roles of hypocrisy and moral emotions in deciding culpability and punishment of criminal and civil moral transgressors. *Cognition and Emotion*, *28*(1), 59–83. https://doi.org/10.1080/02699931.2013.801339

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*(2), 477–480. https://doi.org/10.1016/j.jesp.2010.10.004

Molho, C., Tybur, J., Guler, E., Balliet, D., & Hofmann, W. (2017). Disgust and anger relate to different aggressive responses to moral violations. *Psychological Science*, *28*(5), 609–619. https://doi.org/10.1177/0956797617692000

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. https://doi.org/10.1177/1754073912468165

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

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, *36*(4), 859–866. https://doi.org/10.1111/j.1745-9125.1998.tb01268.x

Pearl, J. (2012). The causal mediation formula—A guide to the assessment of pathways and mechanisms. *Prevention Science*, *13*(4), 426–436. https://doi.org/10.1007/s11121-011-0270-1

Piazza, J., Russell, P. S., & Sousa, P. (2013). Moral emotions and the envisaging of mitigating circumstances for wrongdoing. *Cognition and Emotion*, *27*(4), 707–722. https://doi.org/10.1080/02699931.2012.736859

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. HarperCollins College Division.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

Royzman, E., Atanasov, P., Landy, J. F., Parks, A., & Gepty, A. (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. https://doi.org/10.1037/0022-3514.76.4.574

Russell, P. S., & Giner-Sorolla, R. (2011). Moral anger is more flexible than moral disgust. *Social Psychological and Personality Science*, *2*(4), 360–364. https://doi.org/10.1177/1948550610391678

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. https://doi.org/10.1177/1948550612442913

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

Shen, F. X., Hoffman, M. B., Jones, O. D., Greene, J. D., & Marois, R. (2011). Sorting guilty minds. *New York University Law Review*, *86*(5), 1306–1360.

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. In A. M. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). Taylor & Frances/Routledge.

Skorinko, J. L., Laurent, S., & Bountress, K. (2014). Effects of perspective taking on courtroom decisions. *Journal of Applied Social Psychology*, *44*(4), 303–318. https://doi.org/10.1111/jasp.12222

Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2014). Medflex: An R package for flexible mediation analysis using natural effect models. *Journal of Statistical Software*. Advance online publication. https://doi.org/10.18637/jss.v076.i11

Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, *10*(5), 479–491.

Tetlock, P., Kristel, O., Elson, S., Green, M., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*(5), 853–870. https://doi.org/10.1037/0022-3514.78.5.853

Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. L., Ginther, M. R., Jones, O. D., & Marois, R. (2014). Corticolimbic gating of emotion-driven punishment. *Nature Neuroscience*, *17*(9), 1270–1275. https://doi.org/10.1038/nn.3781

Valeri, L., & Vanderweele, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, *18*(2), 137–150. https://doi.org/10.1037/a0031034

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.

Winterich, K. P., Han, S., & Lerner, J. S. (2010). Now that I'm sad, it's hard to be mad: The role of cognitive appraisals in emotional blunting. *Personality and Social Psychology Bulletin*, *36*(11), 1467–1483. https://doi.org/10.1177/0146167210384710

Received November 29, 2019
Revision received October 13, 2020
Accepted November 4, 2020 ■