Anger and Sadness as Moral Signals

Jason E. Plaks¹, Jeffrey S. Robinson¹, and Rachel Forbes¹

Abstract

Three studies examined the relationship between emotions and moral judgment from an interpersonal perspective. In Studies 1 and 2, participants justified their decisions in sacrificial dilemmas to an imagined interlocutor. Linguistic analyses revealed that Don’t Sacrifice justifications contained more anger-related language than sadness-related language, whereas Sacrifice justifications contained roughly equal proportions of anger and sadness language. In Study 3, participants made character inferences about an actor who chose to/refused to sacrifice one person to save multiple people. We manipulated the actor’s ratio of anger to sadness. Participants rated the Don’t Sacrifice actor more negatively when they displayed high anger relative to sadness but rated the Sacrifice actor negatively whenever they exhibited high anger (independent of sadness). These data highlight novel ways in which actors and observers use emotions to complement the substance of a moral argument.

Keywords

moral judgment, emotion

In recent decades, researchers have made significant discoveries regarding the emotional processes involved in moral judgment (e.g., Cushman et al., 2012; Greene et al., 2004). For example, in the literature on sacrificial dilemmas, researchers have discovered that participants’ level of emotionality positively predicts their tendency to prefer the Don’t Sacrifice option over the Sacrifice option (Greene et al., 2001).¹ Other researchers have found analogous effects with more discrete emotions such as anger (Hutcherson & Gross, 2011) and disgust (Baron et al., 2018).

We suggest that, as helpful as such work has been, researchers in this tradition have typically conceptualized the moral decision maker as an isolated entity. Everyday moral decisions, however, are often embedded in an interpersonal context, in which people must justify their moral decisions and actions to real or imagined observers (e.g., Haidt, 2001; Tetlock, 2002). As such, we suggest that the study of emotion in morality stands to be enhanced by greater scrutiny of emotions’ communicative function.

The present work adopts such an interpersonal approach by asking two questions:

1. Do people spontaneously express distinct emotional profiles to an observer depending on whether they have chosen the Sacrifice or Don’t Sacrifice option?
2. If so, do such distinct emotional profiles systematically influence observers’ character inferences?

The Actor Side of the Equation

Scanlon (1998) proposed that an act should be considered morally acceptable if it can be “justified to others in a way that they cannot reasonably reject.” According to this perspective, moral behavior is inextricably linked with interpersonal communication. Scanlon described this communication as rational, substantive discourse between two engaged individuals.

We suggest that, in practice, people often use rhetorical tools that go beyond the substance of the argument itself. One such tool is emotional expression. Numerous studies indicate that specific emotional displays reliably evoke specific responses from observers. For example, anger typically evokes fear and submission (Dimberg & Ohman, 1996), whereas sadness evokes sympathy and aid (Eisenberg et al., 1989). There is also evidence that specific emotional displays elicit corresponding character inferences (Hareli & Hess, 2010, 2012). Building on these ideas, we examined whether participants would spontaneously complement the substance of their justifications for Don’t Sacrifice and Sacrifice decisions with varying levels of anger- and sadness-related language.

What might be the nature of this variation? Imagine an individual who chooses the Don’t Sacrifice option. Although this individual may express several emotions simultaneously, we hypothesized that the relative proportion would favor anger for two reasons. First, just as those who prefer the Don’t Sacrifice option tend to prefer cognition that is comparatively low in complexity (Moore et al., 2008; Robinson et al., 2019), so too might these people wish to express to their interlocutor that the

¹ University of Toronto, Ontario, Canada

Corresponding Author:

Jason E. Plaks, University of Toronto, 4003 Sidney Smith Hall, 100 St. George Street, Toronto, Ontario, Canada M5S 3G3.

Email: plaks@psych.utoronto.ca
decision not to sacrifice was straightforward and obvious. An expression of anger/frustration may be well-suited to this task (e.g., “Killing is wrong. End of story!”). Indeed, several studies have linked anger (more than other negatively valenced emotions) with expression of certainty (Lerner & Keltner, 2001). In a related vein, Don’t Sacrificers may use anger to assure observers that they are not choosing the Don’t Sacrifice option simply because it is the “do nothing” option, the path of least resistance. Rather, they are acting out of sincere indignation over the idea of using a person as a tool (Kamm, 1996).

Second, deontological tendencies (typically associated with the Don’t Sacrifice option) have been linked with religiosity (McPhetres et al., 2018; Piazza & Sousa, 2014). Religiosity, in turn, has been associated with higher levels of anger toward moral violators (Greer et al., 2005; McGregor et al., 2010). Thus, the deontological emphasis on upholding nonnegotiable rules (e.g., “Thou shalt not kill”) may resonate with some individuals’ religious tendencies and, in turn, increase the accessibility of anger-related responses.

Now imagine an actor who decides to sacrifice one person to save multiple others. To the extent that Sacrificers gravitate toward more complex cognitive strategies (e.g., Patil et al., 2021; Smillie et al., in press), they may show a preference for emotion language that expresses the tragic nature of the no-win situation. Moreover, whereas in typical dilemmas, the Don’t Sacrifice responder’s task is to justify a decision that is the normative default (not killing), the Sacrifice responder must justify killing a person. Expressing sadness may be appropriate in this situation. Individuals who express sadness are perceived to be low in aggression (Hareli & Hess, 2010, 2012) and power (Clarke et al., 1987). Thus, a display of sadness may serve as an appeasement function similar to that of embarrassment (Keltner & Anderson, 2000). More generally, the notion that actors are cognizant of the dispositional inferences that observers might make finds precedent in studies by Rom and Conway (2018), who reported that participants accurately predicted how their dilemma judgments would affect observers’ ratings of their warmth and competence.

The Observer Side of the Equation

Several studies provide evidence that observers often judge Sacrificers to possess lower moral character than Don’t Sacrificers (Everett et al., 2016; Uhlemann et al., 2013). This appears to be due in part to the fact that the decision to sacrifice possesses built-in ambiguity: Is the actor expressing a sincere conviction to utilitarian principles or a dangerous willingness to kill? Robinson et al. (2017) reported, however, that this default suspicion was reduced when the Sacrificer was described as experiencing emotional stress. That is, an expression of sincere conflict allayed concerns that his willingness to kill reflected a dangerous character. Robinson et al. did not, however, identify which particular emotions might accomplish this.

To begin to address this question, in Study 3, we investigated sadness and anger. Based on the reasoning in the previous section, we hypothesized that, for Sacrificers, a display of sadness would be more effective than anger at conveying the inherent tragedy of the situation. Thus, observers would rate a Sacrificer’s character higher when they displayed high (vs. low) sadness. In contrast, we hypothesized that observers would look more favorably upon a Don’t Sacrifice actor when the justification featured a high (vs. low) proportion of anger. This is because a Don’t Sacrifice justification is more likely to include expressions of righteous indignation toward the violation of a cherished rule, which, in turn, may serve to assure the observer that this person is a principled and reliable contributor to the group (Everett et al., 2016; Petit, 2018).

Study 1

In Study 1, we examined whether Sacrificers and Don’t Sacrificers would spontaneously express different levels of anger and sadness in written justifications for their decisions.

Participants

The minimum required sample size to achieve 80% power was determined using G*Power Version 3.1 (Faul et al., 2007). With a 3 (emotion language: anger vs. Sadness vs. anxiety, within-participants) × 2 (decision type: Sacrifice vs. Don’t Sacrifice, between-participants) design, a small effect size (η² = .01), and a moderate correlation between repeated-measures (r = .25), the minimum recommended sample size to achieve 80% power is 240.

We recruited 251 American participants from Amazon’s Mechanical Turk. Due to missing data, repeat participants, and failed attention checks, 28 participants were removed prior to any analyses. The final data set consisted of 223 (140 females, average age = 36.95, SD = 12.41; 56% college degree or higher).

Materials and Procedure

Materials, data sets, analytic plans, and analytic output for all studies are available at: (https://osf.io/8bw4c/). Participants were randomly assigned to one of two dilemmas and were instructed to indicate what they would do in that situation. Next, participants were given 10 min to type a written justification for their choice. Participants then completed a demographics questionnaire and were debriefed.

Dilemmas

Participants were randomly assigned to evaluate either “Nuclear reactor” (Moore et al., 2008) or “Sophie’s choice” (Greene et al., 2008). Each dilemma requires the actor to cause the death of someone in order to yield a net positive outcome (the Sacrifice option) or to not cause the death, but forgo a net positive outcome (the Don’t Sacrifice option; see Online Appendix A for the full text of all dilemmas used in all studies).
Participants’ justifications (mean length = 111.18 words, standard error [SE] = 4.22) were content-analyzed using the Linguistic Inquiry Word Count (LIWC) program (Pennebaker et al., 2015). For each participant’s output, LIWC provided a measurement of the percentage of anger-, sadness-, and anxiety-related words relative to the total word count. (These are the only three specific emotion terms in LIWC2015’s default dictionary.)

We included anxiety-related words in the analyses as a control variable. Circumplex models of affect tend to locate anxiety-related terms (e.g., tense, nervous) in close proximity to anger-related terms, in that both are negative emotions with high arousal (e.g., Remington et al., 2000). Thus, if the pattern for anxiety words tracked that of anger words, we would not be able to conclude that effects involving anger were unique to anger.

**Results**

No analyses were conducted until participant recruitment was completed and all exclusions were performed. Overall, 56% chose the Don’t Sacrifice option in the Nuclear Reactor dilemma, 79% in the Sophie’s Choice dilemma.

Because the LIWC word use data were significantly skewed, we log-transformed the values. To compare the relative percentages of anger, sadness, and anxiety words used in Sacrifice versus Don’t Sacrifice justifications, we conducted a 2 (decision: Sacrifice vs. Don’t Sacrifice) \( \times 3 \) (emotion: anger/sadness/anxiety) mixed-model analysis of variance (ANOVA). This revealed only a significant effect for the interaction term, \( F(2, 460) = 6.96, p < .001, \eta^2_p = .03 \).

Tests of the simple effects revealed that participants justifying a Sacrifice decision used more sadness language than did those justifying a Don’t Sacrifice decision, \( F(1, 230) = 10.79, p < .001, \eta^2_p = .05 \). In contrast, participants justifying a Don’t Sacrifice decision, if anything, used more anger language than did participants justifying a Sacrifice decision, \( F(1, 230) = 3.39, p < .067, \eta^2_p = .02 \). Don’t Sacrificers and Sacrificers used equivalent amounts of anxiety-related language, \( F(1, 230) = 1.39, p = .24 \). Examining the pattern a different way, Don’t Sacrificers used significantly more anger language than sadness language, \( F(1, 159) = 16.62, p < .001, \eta^2_p = .10 \), whereas Sacrificers used an equivalent mix of anger and sadness language, \( F(1, 71) = 2.16, p > .14, \eta^2_p = .03 \) (see Figure 1). In addition, as a reality check, we subsequently asked two human raters to rate the levels of anger, sadness, and anxiety present in each participant’s text. These results, which are described in Online Appendix B, paralleled the LIWC results.

**Study 2**

The primary aim of Study 2 was to replicate the effect with three new dilemmas.

**Participants**

The minimum required sample size to achieve 80% power was determined using G*Power Version 3.1 (ANOVA: repeated measures, within-between interaction). Using the effect size for
the interaction observed in Study 1 ($R^2 = .02$) and a moderate correlation between repeated-measures ($r = .25$), the minimum recommended sample size to achieve 80% power was 60 participants. A total of 128 undergraduate students in an introductory psychology course enrolled in the study. Eleven participants did not provide data. The final set included 117 participants (95 females, average age $= 18.42, SD = 1.80$).

**Materials and Procedure**

**Procedure**

Upon being individually seated at a laboratory computer, participants listened to an audio recording of a narrator describing three dilemmas. After each dilemma, participants were instructed to say aloud into a microphone what they would do in that situation. They were then instructed to imagine that they were standing face-to-face with someone who disagreed with their decision. Their task was to convince that person that their own decision was morally correct. This process was repeated for all three dilemmas. (Thus, “Dilemma” was a within-participant variable, unlike in Study 1.) At the conclusion, they completed a demographics survey and were debriefed.

**Dilemmas**

Participants read three commonly used sacrificial dilemmas (“Doctor,” “Researcher,” “Commander in chief”; Greene et al., 2008; Robinson et al., 2015).

**Justifications**

Spoken justifications were transcribed and analyzed using LIWC (Pennebaker et al., 2015); mean word count = 155.76, $SE = 7.30$. As in Study 1, we included anxiety words in the analysis to test whether the pattern predicted for anger would be limited to anger and not another high arousal, negative emotion.

**Results**

Participants generally showed a slight preference for the Sacrifice option (“Commander in chief” = 60%, “Doctor” = 62.5%, “Researcher” = 62.5%). To compare the relative proportions of anger, sadness, and anxiety language that participants used as a function of their moral decision, we conducted a two-level multilevel model, using the lmer function from the lme4 package (Bates et al., 2015) in R3.4.2 (R Core Team, 2017). Multilevel modeling was appropriate here because, unlike in Study 1, all participants judged three separate scenarios. Thus, the proportion of anger, sadness, and anxiety words was nested within each participant’s decision on each dilemma. Because the interaction between type of emotion (Level 1) and dilemma decision (Level 2) was measured at different levels, we modeled a random slope for the Level 1 predictor (emotion), in addition to the random intercept.

This analysis revealed a marginal effect of participants’ decision, $F(1, 670.63) = 2.93, p = .09, R^2_g = .004$, a significant effect of emotion, $F(2, 301.74) = 18.44, p < .001, R^2_g = .11$, and, akin
to Study 1, a significant interaction, \(F(2, 609.45) = 5.73, p < .01, R^2 = .02.\)

Simple effects tests revealed that participants justifying a Don’t Sacrifice decision used marginally more anger language than did participants justifying a Sacrifice decision, \(b = .20, t(718.16) = 1.91, p = .06\). At the same time, Sacrifice responders used significantly more sadness language than did Don’t Sacrifice responders, \(b = .33, t(587.32) = 3.32, p < .001\). Examining the effect a different way, Don’t Sacrifice justifications were characterized by higher use of anger language than sadness language, \(b = .75, t(305.45) = 6.34, p < .001\). In contrast, Sacrifice justifications were characterized by closer-to-equal use of anger and sadness language, \(b = .22, t(193.15) = 2.37, p < .05\). Don’t Sacrifice and Sacrifice responders did not differ in their use of anxiety-related words, \(b = -.11, t(618.95) = -1.01, p = .28\). In sum, these results, using a different set of dilemmas, closely parallel those of Study 1 (see Figure 2).

**Study 3**

Based on Studies 1 and 2, one might expect observers to rate the character of Don’t Sacrificers more positively to the extent that they display predominantly anger and Sacrificers more positively to the extent that they display predominantly sadness. If so, this would suggest that actors generally possess accurate knowledge of how to “pull the strings” of observers (Rom & Conway, 2018). However, if a different pattern were to emerge (e.g., a display of anger lowers observers’ opinion of Don’t Sacrificers), this would suggest that actors have imperfect knowledge of the effects their emotional signals have on observers. To investigate this question, we manipulated actors’ levels of anger and sadness to assess the independent effects of each emotion on observers’ judgments of the actor.

**Method**

**Participants**

To provide an estimate of the minimum sample size required to achieve 80% power, we conducted a power analysis using the ANOVA calculator in G*Power (Faul et al., 2007). The main hypothesis required testing the interaction between the emotion manipulation and the fictitious participant’s moral judgment. We assumed a small effect size \((f = .10)\). We specified a Type I error rate of .05, 80% power, eight groups (actor anger: high vs. low \(\times\) actor sadness: high vs. low \(\times\) actor decision: Sacrifice vs. Don’t Sacrifice), and a small correlation between the repeated measures \((r = .20)\). This test revealed a minimum sample size of 680.

Participants were recruited via Mechanical Turk. From an initial total of 1,115 individuals, 57 were eliminated prior to any analyses because they did not successfully meet the inclusion criteria (e.g., failed attention checks, multiple attempts, improbable speed), leaving a final sample of 1,058 (573 females; average age = 37.89; 59.52% with a college degree).

**Materials and Procedure**

**Procedure.** Participants were instructed that they would read three dilemmas and would be asked to indicate what they would do in the situation. The cover story stated that previous participants had evaluated the same scenarios while having their brain activity recorded. The instructions led participants to believe that the measure of brain activity we used (electroencephalography; EEG) is able to distinguish between the type and intensity of discrete emotions.

We recognized that the concurrent emotional experience (ostensibly) revealed by neural activity differs in many ways from the post hoc verbal justifications measured in Studies 1 and 2. Nonetheless, we chose EEG as an index of the actor’s experienced emotion due to pervasive assumptions that neural activity represents an accurate, unfakeable pipeline into mental processes (e.g., Weisberg et al., 2008). In the present study, we wished participants to believe that the actor’s emotions were genuine (i.e., unaffected by self-presentational concerns) because our aim was to measure participants’ attributions regarding the emotions themselves, not the act of strategic emotional expression.

Participants were shown a random previous participant’s three decisions, accompanied by a bar graph indicating the emotions that they had experienced during each decision. We systematically manipulated the amount of anger and sadness that the actor displayed. Anxiety, disgust, and compassion were kept constant (near the midpoint; see Figure 3). This manipulation followed a full 2 (actor anger: high vs. low) \(\times\) 2 (actor sadness: high vs. low) between-subjects design because doing so would allow us to identify particular combinations of high versus low anger and sadness that impress versus “turn off” observers.

After viewing the actor’s choice (Sacrifice vs. Don’t Sacrifice) and emotion levels, participants rated the actor’s moral character. This process repeated three times. The actor’s emotional profile was kept constant across the three scenarios. At the conclusion, participants completed a demographics questionnaire and were debriefed.

**Dilemmas.** The computerized program randomly selected from a set of nine dilemmas. Three personal dilemmas were scenarios that have been widely used in previous work (“Footbridge,” Greene et al., 2001; “Bike Week” and “Vaccine Test,” Moore et al., 2008). In addition, we included three self-sacrifice dilemmas and three impartiality dilemmas that Kahane and colleagues (2015, 2018) have argued are better suited to measure key features of utilitarian philosophy. All dilemmas were worded in the first person. After reading each dilemma, participants indicated what they would do via a dichotomous choice question.

**Moral character judgments.** Participants rated the actor by indicating their level of agreement (1 = strongly disagree; 7 = strongly agree) with six statements adapted from (Critcher et al., 2013) and (Uhlmann et al., 2013; e.g., “I think this person
is extremely ethical."). These statements showed high internal reliability across dilemma types (personal: $\alpha = .97$, self-sacrifice: $\alpha = .97$, and impartial: $\alpha = .94$) and were thus aggregated into a single moral character index (maximum score = 42).

**Results**

To measure the combined effects of actors’ decisions and levels of anger/sadness on observers’ character judgments, we conducted a 2 (actor anger: high vs. low) $\times$ 2 (actor sadness: high vs. low) $\times$ 2 (actor decision: Sacrifice vs. Don’t Sacrifice) between-participants ANOVA.

This analysis revealed main effects for actor anger, $F(1, 1040) = 18.40, p < .001, \eta^2_p = .02$, actor sadness, $F(1, 1040) = 13.85, p < .001, \eta^2_p = .01$, and actor decision (marginal), $F(1, 1040) = 2.73, p = .08, \eta^2_p = .003$, as well as an actor anger $\times$ actor sadness interaction, $F(1, 1040) = 4.10, p < .05, \eta^2_p = .004$. The three-way interaction did not reach significance, $F(1, 1040) = 2.07, p = .15$.

To better understand the pattern, we conducted simple effects comparisons between cells of interest. These revealed, first, that participants considered the actor who made the Don’t Sacrifice decision accompanied by high anger/high sadness more moral than the actor who made the same decision accompanied by high anger/low sadness, $F(1, 1040) = 9.65, p = .002$. This difference was not evident, however, when the actor made the Sacrifice decision, $F(1, 1040) = 2.72, n.s.$

As depicted in Figure 4, in three of the four emotion conditions, participants judged the Sacrifice and Don’t Sacrifice actor to be equivalent. The lone exception was the high anger/high sadness condition. In this condition, the Don’t Sacrifice actor was rated more positively than the Sacrifice actor, $t(1040) = 6.11, p < .02$.

**Did Participants’ Own Decision Moderate the Effect?**

Recall that participants viewed the decisions and emotions of three actors. In each case, participants were also asked to indicate their own decision (Sacrifice or Don’t Sacrifice). Thus, for analytic purposes, participants’ own decisions were aggregated into a continuous variable (own decision) that was scored as follows: $-1 = $ all Don’t Sacrifice; $-.50 = $ 2 Don’t Sacrifice/1 Sacrifice; $.50 = $ 1 Don’t Sacrifice/2 Sacrifice; $1 = $ all Sacrifice. When this variable was added to the model, it did not contribute any significant main effects or interactions, all $F$s $< 1.61$, while the effects reported above remained significant.

In other words, even participants who themselves preferred the Sacrifice option found the high anger/high sadness Sacrifice actor to possess a worse character than the high anger/high sadness Don’t Sacrifice actor. This finding is consistent with prior work in which participants’ own preferences did not influence their judgments of the character of Sacrificer vs. Don’t Sacrificer actors (Everett et al., 2016; Robinson et al., 2017; Uhlmann et al., 2013; cf. Bostyn & Roets, 2017). For example, Uhlmann et al. (2013) reported that participants judged
throwing a dying man out of a lifeboat to prevent it from sinking (thereby saving several other passengers) as the correct choice but also judged the thrower to possess a more negative character than someone who would not throw the man overboard.

**Summary**

The pattern in Study 3 suggests a more nuanced phenomenon than we had hypothesized. The difference in observers’ judgments of Sacrifice and Don’t Sacrifice actors concerned the relative proportion of anger to sadness. For Don’t Sacrificers, observers weighed both anger and sadness in their character judgment. For Sacrificers, observers weighed only anger and largely ignored sadness. In the General Discussion, we explore potential reasons for this dissociation.

This pattern also suggests that actors often mis-predict how their emotions will be interpreted by observers. Based on the results of Studies 1 and 2, one might expect observers to rate a high anger Don’t Sacrificer more positively than a low anger Don’t Sacrificer. Instead, observers generally took a dim view of both high-anger Don’t Sacrificers and high-anger Sacrificers, even if they agreed with the decision. Moreover, based on Studies 1 and 2, one might expect that high sadness would raise observers’ opinion of Sacrificers. Instead, high sadness raised observers’ opinion of Don’t Sacrificers but had little effect on their opinion of Sacrificers.

**General Discussion**

These studies investigated two questions: (1) Do actors spontaneously express distinct emotional profiles when justifying Sacrifice versus Don’t Sacrifice decisions? and (2) Do observers incorporate these emotional signals into their assessments of the actor’s character?

Studies 1 and 2 suggest that the answer to Question #1 is “Yes.” Don’t Sacrifice justifications were accompanied by a higher relative proportion of anger to sadness, whereas Sacrifice justifications featured equivalent proportions of anger and sadness. Why might actors pull on these different emotional strings? We suggest that the Sacrifice decision requires the actor to assure observers that although they had decided that it was proper to kill someone, this decision was not driven by a dangerous character. Showing sadness may be one way to accomplish this. On the other hand, deciding not to sacrifice—and thereby forfeiting an outcome that, in fact, benefits more people—may encourage the actor to emphasize language that expresses the inviolability of a clear, simple rule. Anger may prove effective in this regard.

The data from Study 3 suggest that the answer to Question #2 is a qualified “Yes.” Contrary to our hypothesis, observers did not rate a high anger Don’t Sacrificer more positively than a low anger Don’t Sacrificer. Instead, observers rated a Don’t Sacrificer who displayed high anger paired with high sadness more favorably than a Don’t Sacrificer who displayed high anger and low sadness. That is, the Don’t Sacrifice actor’s sadness “softened” the negative effect of his or her anger. In contrast, the Sacrificer’s sadness did not soften the negative effect of his or her anger.

Why do observers’ judgments of Sacrificers’ character focus mostly on anger and largely ignore sadness? One possibility is that observers’ default assumption is to be skeptical of a Sacrificer’s character (Kreps & Monin, 2014). As such, observers may be less inclined to give a Sacrificer the benefit...
of the doubt. Thus, the mere presence of high anger may represent additional, vivid evidence (beyond the act of killing) that the Sacrificer is a dangerous person.

A second possibility is that observers, having witnessed the actor make a sacrifice decision, must determine whether this decision to kill was morally motivated. Because killing involves the violation of a widely accepted rule, observers likely expect the Sacrificer to convey some degree of uncertainty or humility. Anger, however, has been linked with certainty and assertiveness (Trope, 1986), thus violating this expectation of humility. Further research is needed to test these possibilities more directly.

**Questions, Limitations, and Future Directions**

Did participants in Studies 1 and 2 express their true emotions or did they “dial up”/“dial down” their anger and sadness for self-presentational purposes? For two reasons we have opted to describe participants’ emotions as a motivated (though likely unconscious) display. First, the motivated model fits well with the idea of emotions’ communicative function (Keltner & Anderson, 2000), a concept that informs the present theorizing. Second, LIWC, as a measure of linguistic content, by definition, does not permit inferences regarding mental experience in the absence of language. Nonetheless, we recognize that the present data provide little direct evidence for a motivated, strategic account. Future researchers may consider using methods from the literature on motivated cognition (e.g., Molden & Higgins, 2005) to assess the role of motivation more directly.

Why didn’t Sacrificers display equivalent or even higher levels of anger than Don’t Sacrificers? After all, Sacrifice responders may have felt particularly compelled to express anger given that (a) they were choosing the position that invites greater uncertainty about their motives and (b) anger conveys certainty (Trope, 1986). The data indicate that, in fact, Sacrificers did display considerable anger—more anger, in fact, than sadness in Study 2. At the same time, Sacrificers (more so than Don’t Sacrificers) simultaneously lessened their anger with sadness. Taken together, these data suggest that Sacrificers are more likely to express both anger and sadness simultaneously.

**Emodiversity**

Much of the literature on sacrificial dilemmas has characterized Sacrificers as generally low in emotion (Greene et al., 2001; Koenigs et al., 2012). Studies 1 and 2 indicate that, at least when it comes to verbal expressions, Sacrificers do not express less emotion as much as they express more evenly distributed emotion. This suggests that Sacrificers place higher value on “emodiversity,” a trait associated with wisdom (Grossmann et al., 2019). In other words, Sacrificers (more than Don’t Sacrificers) may wish to express multiple, competing emotions, possibly to convey epistemic humility. We encourage future researchers to explore this idea more directly.

Participants in all studies were recruited from a western context. Given the evidence of cultural differences in both moral values (Fiske & Rai, 2015; Haidt & Joseph, 2004) and emotional expression (Mesquita & Boiger, 2014), the types of emotions displayed during moral justifications are likely to vary by culture. In addition, the lay theories that individuals hold regarding the link between emotional experience and moral character are also likely to vary by culture. Thus, we encourage future researchers to attempt to replicate these studies across multiple cultures.

It is also important to note that although LIWC is an oft-used measure of linguistic content, machine learning–based approaches offer numerous advantages, including a greater ability to analyze the context in which each word is used (e.g., Iliev, Dehghani, & Sagi, 2014). We encourage future researchers to extend our findings using these more sensitive techniques.

**Conclusion**

Studies 1 and 2 suggest that people complement the verbal content of their moral justifications with specific emotional combinations, in an attempt to convey positive moral character. The Study 3 data go on to suggest that observers do generate distinct moral character inferences from the actors’ emotions—but not necessarily the inferences that actors expect. Taken together, these studies represent a starting point for future research on strategies and pitfalls of moral communication.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**ORCID iD**

Jason E. Plaks @ https://orcid.org/0000-0003-2706-8067

**Supplemental Material**

The supplemental material is available in the online version of the article.

**Notes**

1. We use the terms Don’t Sacrifice and Sacrifice due to criticisms regarding the appropriateness of using the terms “Utilitarian” and “Deontological” to characterize laypeople’s judgment (e.g., Kahane et al., 2018).

2. Principled Utilitarians may express anger toward someone who prefers the Don’t Sacrifice option (“Five is more than one!”). Most people who choose the Sacrifice option, however, do not qualify as principled Utilitarians (Conway et al., 2018).]

3. In most dilemmas, the kill:save ratio is low enough to place most participants in the horns of a dilemma. There are, however, variables that reduce the need to justify killing one person, including
increasing the number of saved individuals (e.g., Tremoliere & Bonnefon, 2014).

4. Additional analyses that included the variable dilemma did not yield any significant effects in any of the studies.

5. As a reality check, human coders coded participants’ text for anger, sadness, and anxiety. As described in Online Appendix B, the results closely paralleled the Linguistic Inquiry Word Count output.

References

Baron, J., Gurcay, B., & Luce, M. F. (2018). Correlations of trait and state emotions with utilitarian moral judgments. *Cognition and Emotion, 32*(1), 116–129.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bostyn, D. H., & Roets, A. (2017). Trust, trolleys and social dilemmas: A replication study. *Journal of Experimental Psychology: General, 146*(5), e1–e7.

Clarke, M. S., Ouellette, R., Powell, M. C., & Millberg, S. (1987). Recipient’s mood, relationship type, and helping. *Journal of Personality and Social Psychology, 53*(1), 94–103.

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition, 179*, 241–265.

Crichter, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*(3), 308–315.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion, 12*(1), 2–7.

Dimberg, U., & Öhman, A. (1996). Behold the wrath: Psychophysiological responses to facial stimuli. *Motivation and Emotion, 20*, 149–182.

Eisenberg, N., Fabes, R. A., Miller, P. A., Fultz, J., Shell, R., Mathy, R. M., & Reno, R. R. (1989). Relation of sympathy and personal distress to prosocial behavior: A multimethod study. *Journal of Personality and Social Psychology, 57*, 45–66.

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772–787.

Faulk, F., Erdfeider, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Fiske, A. P., & Rai, T. S. (2015). *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Haidt, J., Morelli, S. A., and Oakes, H., & Santos, H. C. (2019). Wise reasoning benefits from emodiversity, irrespective of emotional intensity. *Journal of Experimental Psychology: General, 148*(5), 805–823.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133*, 55–66.

Hareli, S., & Hess, U. (2010). What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception. *Cognition and Emotion, 24*(1), 128–140.

Hareli, S., & Hess, U. (2012). The social signal value of emotions. *Cognition and Emotion, 26*(3), 385–389.

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*, 719–737.

Iliev, R., Dégagné, M., & Sagi, E. (2014). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition, 7*, 265–290.

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*(2), 131–164.

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). “Utilitarian” judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209.

Kallmann, F. M. (1996). *Morality, mortality, Vol 2: Rights, duties, and status*. Oxford University Press.

Keltner, D., & Anderson, C. (2000). Saving face for Darwin: The function and uses of embarrassment. *Current Directions in Psychological Science, 9*, 187–192.

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social, Cognitive, and Affective Neuroscience, 7*, 708–714.

Kreps, T. A., & Monin, B. (2014). Core values versus common sense views appear less rooted in morality. *Personality and Social Psychology Bulletin, 40*(11), 1529–1542.

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology, 81*(1), 146–159.

McGregor, I., Nash, K., & Prentice, M. (2010). Reactive approach motivation (RAM) for religion. *Journal of Personality and Social Psychology, 99*, 148–161.

McPhetres, J., Conway, P., Hughes, J. S., & Zuckerman, M. (2018). Reflecting on god’s will: Reflective processing contributes to religious peoples’ deontological dilemma responses. *Journal of Experimental Social Psychology, 79*, 301–314.

Mesquita, B., & Boiger, M. (2014). Emotions in context: A socio-behavioral, and biomedical sciences. *Behavior Research Methods, 46*(1), 148–161.

Molden, D. C., & Higgins, E. T. (2005). Motivated thinking. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of...*
thinking and reasoning (pp. 295–317). Cambridge University Press.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. Psychological Science, 19(6), 549–557.

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., Silani, G., Cikara, M., & Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. Journal of Personality and Social Psychology, 120(2), 443–460.

Pennebaker, J. W., Booth, R. J., Boyd, R., & Francis, M. E. (2015). Linguistic inquiry and word count: LIWC2015. Pennebaker Conglomerates.

Petit, P. (2018). The birth of ethics: Reconstructing the role and nature of morality. Oxford University Press.

Piazza, J., & Sousa, P. (2014). Religiosity, political orientation, and consequentialist moral thinking. Social Psychological and Personality Science, 5, 334–342.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.r-project.org/

Remington, N. A., Fabrigar, L. R., & Visser, P. S. (2000). Reexamining the circumplex model of affect. Journal of Personality and Social Psychology, 79(2), 286–300.

Robinson, J. S., Joel, S., & Plaks, J. E. (2015). Empathy for the group versus indifference toward the victim: Effects of anxious and avoidant attachment on moral judgment. Journal of Experimental Social Psychology, 56, 139–152.

Robinson, J. S., Page-Gould, E., & Plaks, J. E. (2017). I appreciate your effort: Asymmetric effects of actors’ exertion on observers’ consequentialist versus deontological judgments. Journal of Experimental Social Psychology, 73, 50–64.

Robinson, J. S., Xu, X., & Plaks, J. E. (2019). Disgust and deontology: Trait sensitivity to contamination promotes a preference for order, hierarchy, and rule-based moral judgment. Social Psychological and Personality Science, 10(1), 3–14.

Rom, S., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. Journal of Experimental Social Psychology, 74, 24–37.

Scanlon, T. M. (1998). What we owe to each other. Harvard Belknap Press.

Smillie, L. D., Katic, M., & Laham, S. M. (2021). Personality and moral judgment: Curious consequentialists and polite deontologists. Journal of Personality, 89, 549–564.

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. Psychological Review, 109(3), 451–471.

Tremoliere, B., & Bonnefon, J-F. (2014). Efficient kill-save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. Personality and Social Psychology Bulletin, 40, 923–930.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. Psychological Review, 93(3), 239–257.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. Cognition, 126, 326–334.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. Journal of Cognitive Neuroscience, 20, 470–477.

Author Biographies

Jason E. Plaks is an associate professor in the Department of Psychology at the University of Toronto. His research examines the relationships between motivation and social cognition.

Jeffrey S. Robinson is a researcher in the Department of Psychology at the University of Toronto. His research examines the psychological processes involved in moral judgment.

Rachel Forbes is a Ph.D candidate in Social/Personality Psychology. Her primary research interests focus on morality, specifically individuals’ perceptions of others’ unethical behavior. She is interested in how moral outrage is influenced by social relational factors like power or closeness. She also investigates whether and how individuals can be seen as redeemed after acting unethically.

Handling Editor: Igor Grossmann