Reframing of moral dilemmas reveals an unexpected “positivity bias” in updating and attributions

Minjae J. Kim1*, Jordan Theriault2*, Joshua Hirschfeld-Kroen1, Liane Young1

*Equal contributors

1 Department of Psychology and Neuroscience, Boston College, Chestnut Hill, MA
2 Department of Psychology, Northeastern University, Boston MA

Corresponding author at: minjae.kim@bc.edu

Abstract: People often encounter new information about others that may require them to dynamically update their moral judgments about them. While prior work on impression updating has examined the process of moral updating following the observation of additional moral behaviors that violate prior expectations, relatively little is known about how moral updating occurs when new contextual information directly reframes prior behaviors, casting them as more or less moral than they were on an initial interpretation. The present work sought to identify whether such moral reframing is differentially sensitive to the qualitative character of reframing information. Study 1 found that the negativity bias, well-documented by prior work on moral impression updating, can be reversed, such that immoral-to-moral reframing results in more updating than moral-to-immoral reframing. Study 2 replicated this pattern and found an additional asymmetry: moral reframing information was interpreted as providing more situational (vs dispositional) information, compared to immoral reframing information. Taken together, the present findings show that theories of moral impression updating should consider distinctions between the addition of new moral information and the reframing of prior information, as well as the impact of reframing on attributions.

Keywords: moral updating, negativity bias, situational/dispositional attributions
1. Introduction

People form moral impressions about others based on their behaviors and dynamically update these impressions in light of new information. A large body of social psychological research has identified factors that influence moral updating in the face of counter-attitudinal information. One prominent line of work has identified a *negativity bias* in moral updating, such that initial and updated moral judgments depend more on negative than positive moral information (Skowronski & Carlston, 1989; Mende-Siedlecki, Baron, & Todorov, 2013; Cone & Ferguson, 2015). This work has typically relied on presenting participants with sequences of decontextualized behaviors (e.g., “Laura heckled a woman speaking on human rights; Laura translated items for a foreigner in a restaurant; Laura gave money to charity”, Mende-Siedlecki & Todorov, 2016). By contrast, relatively little work has examined how moral updating unfolds when new information comes to light that provides additional context, or a *reframing*, of prior behavior. Earlier work in this area has shown that **reframing** initially **immoral** behaviors can reverse both implicit and explicit negative first impressions (Mann & Ferguson, 2015). To our knowledge, however, there been no empirical test of whether there is **more** updating when moral behaviors are reframed as immoral, relative to the reverse—reframing immoral behaviors as moral. Research on negativity bias in moral updating leads to a clear prediction that moral to immoral reframing should be stronger, and, given this, the present research set out to test whether negativity bias actually does persist in the context of moral reframing.

1.1. Negativity bias in updating

Importantly, negativity bias is neither a homogeneous nor a universal phenomenon. Several lines of research on impression updating have revealed both 1) properties of negative behaviors (beyond valence) that explain their dominance for updating, and 2) properties of
behaviors that explain differences in updating, in the absence of a negativity bias. One such property is perceived behavioral frequency. Mende-Siedlecki and colleagues (Mende-Siedlecki, Baron, & Todorov, 2013) found that immoral behaviors and highly competent behaviors elicit greater impression updates (compared to moral behaviors and incompetent behaviors) because they are understood to be less frequent than their counterparts, in line with the idea that behaviors that are understood to be more rare are more diagnostic of true character. Relatedly, measures of diagnosticity can also explain differences in updating. In a study using implicit character evaluations, Cone and Ferguson (2015) found that an extremely negative behavior (mutilating a small animal) moved initially positive evaluations more than an extremely positive behavior (donating a kidney) moved initially negative evaluations. Importantly, the more participants viewed the negative action as offensive, the more they viewed the action as reflective of true character, and these perceptions of diagnosticity, in turn, predicted the degree of implicit impression updating (although see Newman, Bloom, & Knobe, 2014 for evidence that people see others’ true selves as fundamentally good).

Other work has shown that the relevance of a behavior for inferring interpersonal intent also matters for updating. Brambilla and colleagues (2019) found that morally relevant behaviors elicit greater impression updates compared to behaviors in the domains of competence or sociality, because they are seen as containing more information about interpersonal intentions. It should be noted that, in this paradigm, no valence asymmetry was observed when comparing impression updates following moral versus immoral behaviors, and mediational analyses did not support a behavioral frequency account of differences in updating.

1.2. Reversing impressions through reinterpretation
Much of the extant work on impression updating typically presents a sequence of disconnected facts about a person or a sequence of disparate behaviors they have performed, e.g., some number of moral behaviors followed by some number of immoral behaviors (Mende-Siedlecki, Cai, & Todorov, 2013; Brambilla, Carraro, Castelli & Sacchi, 2019; Gawronski, Rydell, Vervliet & De Houwer, 2010; Petty, Tormala, Briñol & Jarvis, 2006). In reality, information is rarely decontextualized—for example, when we learn that someone has done something wrong, we often also learn the context for the behavior that may ultimately provide some justification or at least explanation. Indeed, we may be persuaded to question others’ motives, or to reconsider a behavior that at first appeared inexcusable. Mann and Ferguson (2015) illustrate such a case, of a Nazi who employed Jewish workers as a cheap source of labor at the dawn of World War II. On its own, this information is enough to elicit powerful feelings of moral outrage. However, it is then revealed that this man—Oskar Schindler—went to great lengths to keep his Jewish workers from being killed, even risking his own life to protect them. For most, this new information elicits a dramatic update in their moral judgments about Schindler. Indeed, Mann and Ferguson’s studies (2015) show that negative implicit evaluations can be undone if observers are given additional details that reframe seemingly bad prior behavior in a positive light. Crucially, the negative implicit evaluation is reversed only when the new positive information can be used to reinterpret bad prior behavior, and not when the new positive information is seen as irrelevant to the prior behavior. This evidence suggests that a key component of (implicit) impression change is the relationship between new and old information: specifically, whether the new information affords a reinterpretation of past events.

In sum, prior work indicates that behavioral frequency, diagnosticity, intention information, and reinterpretation can all affect impression updating. The present work seeks to
extend and refine our understanding of the relationship between reinterpretation and updating, which may be particularly relevant to real-life moral judgment. Providing people with a moral reinterpretation of initially immoral behaviors has been shown to be a robust method for reversing both implicit and explicit negative first impressions (Mann & Ferguson, 2015). To our knowledge, however, there is no research exploring whether the degree of moral updating changes when people are given the exact same information in reverse order, such that moral behaviors get reframed in an immoral light, and immoral behaviors get reframed in a moral light. Is there more updating when moral information is reframed as immoral, relative to the reverse, as research on negativity biases in moral updating would suggest? And, if a valence asymmetry is found, what is the informational feature underlying the asymmetry?

1.3. Reframing tragic and taboo dilemmas

We explored these questions across two experiments examining explicit moral judgments following both positive and negative reframing information. Specifically, we first presented scenarios framing someone’s behavior as either moral or immoral, and later presented additional information intended to reframe that behavior as the opposite. To design scenarios that could test the relative degree of updating in response to moral-to-immoral and immoral-to-moral reframing, we relied on the rich framework of tragic and taboo dilemmas (Fiske & Tetlock, 1997; Tetlock et al., 2000; Tetlock, 2002; 2003; McGraw & Tetlock, 2005; Ginges, Atran, Medin, & Shikaki, 2007; Gregory & Irwin, 2007; Bartels, 2008; Hanselmann & Tanner, 2008; Lichtenstein, Mandel, & Vartanian, 2008; Tannenbaum, Uhlmann, & Diermeier, 2011). The examples used in this study are simple, but the framework of tragic and taboo dilemmas has been applied across a variety of contexts with practical relevance to important issues (including negotiations for peace in the Israel-Palestine conflict; Ginges et al., 2007). In tragic dilemmas, decision-makers are
faced with a choice between two competing moral values, such that they must choose the “lesser of two evils”, while sacred values (sometimes others’ lives and livelihoods) hang in the balance. Imagine a scenario used in the present work: a fishing boat captain, Jeremy, must choose whether to buy expensive new fishing nets that would kill fewer dolphins, where, to compensate for the expense, he would also need to lay off many of his crew (who would not find new jobs easily). Critically, the ambiguity of these tragic dilemmas makes moral judgment difficult, as it is unclear which choice is “morally correct” (Tetlock et al., 2000). Given this, if Jeremy decides not to buy fishing nets, sacrificing the lives of dolphins, then people may still judge him as moral (or at least not immoral) because his decision ultimately protected the livelihoods of his workers.

By comparison, taboo dilemmas present individuals with a choice between a moral concern and economic self-interest. Imagine another fisherman, Steve, who runs a side business selling dolphin fins on the black market. When Steve learns about the new fishing nets, he must choose between saving dolphins and keeping his side-business. Most people believe that prioritizing economic self-interest over the lives of others is wrong (Tetlock, 2003), and, as such, people may judge Steve as immoral if he chooses to maximize his profit at the expense of the dolphins.

Critically, tragic and taboo cases can be reframed (Atran & Axelrod, 2008). Now imagine that Jeremy and Steve from the previous scenarios are the same person: Gregory. Initially, when Gregory favored his workers’ jobs over the lives of the dolphins, he appeared to be making a moral decision in the context of a tragic dilemma. However, his decision can be reframed as an immoral decision in a taboo dilemma by revealing his side business. The information can also be presented in reverse: first establishing an immoral decision in a taboo dilemma, then elaborating
on how Gregory’s workers are actually helped by his decision, suggesting he made a moral decision in a tragic dilemma.

1.4. Present Work

We used tragic and taboo dilemmas to explore possible asymmetries in moral updating. The stimuli take two forms: Moral-to-Immoral reframing, where agents’ decisions begin as moral (in the context of tragic dilemmas) and are reframed as immoral (in the context of taboo dilemmas), and Immoral-to-Moral reframing, where agents’ decisions begin as immoral (taboo) and are reframed as moral (tragic) (Figure 1). In either form, participants report a first-pass moral judgment, and then report a second-pass moral judgment made after reframing, with moral updating calculated as the difference between the two judgments. As prior work has identified asymmetries in moral judgments about agents’ actions (act-based judgments) and moral judgments about agents themselves (person-based judgments) (Uhlmann, Zhu, & Tannenbaum, 2013), we measured moral updating for both types of judgments to test the robustness of any effects that were observed.

In Study 1, consistent with prior work in impression formation, we predicted that we would observe a negativity bias, (Mende-Siedlecki, Cai, & Todorov, 2013; Baumeister et al., 2001; Rozin & Royzman, 2001), such that Immoral-to-Moral reframing would elicit less updating than Moral-to-Immoral reframing. Contrary to this prediction, we observed the opposite pattern in Study 1: updating was greater for Immoral-to-Moral reframing than for Moral-to-Immoral reframing. We hypothesized that this discrepancy may be due to an asymmetry in the interpretation of immoral and moral information—i.e., they may differently afford dispositional and situational attributions (Fein, 1996; Fein et al., 1990; Gilbert & Jones, 1986; Gilbert & Malone, 1995; for review see Gawronski, 2004). Study 2 directly investigated this asymmetry in
dispositional and situational attributions, revealing that moral information (both when presented initially and when presented as reframing information) was interpreted as providing more information about the situation than about the agent’s disposition, whereas there was no such difference in attributions for immoral information. Furthermore, the more participants interpreted reframing information as providing situational information, the more likely they were to engage in positive moral updating.

2. Study 1

Study 1 examined asymmetries in moral updating between Moral-to-Immoral and Immoral-to-Moral reframed scenarios. We predicted that Moral-to-Immoral updating would be greater than Immoral-to-Moral updating, consistent with prior work on negativity bias (Mende-Siedlecki, Cai, & Todorov, 2013; Baumeister et al., 2001; Rozin & Royzman, 2001).

2.1. Methods and Materials

2.1.1. Participants

Participants were recruited through Amazon Mechanical Turk in exchange for payment. The final sample consisted of 122 adults (63 female; $M_{\text{Age}} = 33.25$ years, $SD_{\text{Age}} = 11.17$ years), after excluding 8 participants for failing a simple attention check, and 17 participants for quitting before completing the survey. The Boston College Institutional Review Board approved the study, and each participant provided informed consent before beginning the survey.

2.1.2. Stimuli

We created 24 detailed moral scenarios, either designing novel scenarios or adapting them from prior work (Critcher et al., 2012; Tetlock et al., 2000; Uhlmann, Zhu & Tannenbaum, 2013). Each scenario initially presented an agent’s moral or immoral decision given a dilemma (Initial Condition), and was supplemented with either subsequent reframing or control
information (*Updating Condition*), designed to reverse the initial moral judgment (reframing), or to be morally irrelevant (control) (Figure 1). Participants read each scenario and made moral judgments at two time points (*Time Point*): a first-pass judgment, made when the agent’s moral or immoral decision was presented, and a second-pass judgment, made after the scenario was reframed (or after control information was presented). Moral updating was calculated as the difference between second-pass and first-pass moral judgments. The combination of Time Point, Initial Condition, and Updating Condition created six conditions: (a) Initial Moral, (b) Initial Immoral, (c) Moral-to-Immoral, (d) Immoral-to-Moral, (e) Moral–Control, (f) Immoral–Control. Critically, each scenario was composed of segments, which could be rearranged to generate the conditions of interest. Rearranging these segments ensured that at the end of both Moral-to-Immoral and Immoral-to-Moral scenarios, participants had received the exact same information, only presented in different orders (Figure 1). Each participant read 24 scenarios (6 Moral-to-Immoral, 6 Immoral-to-Moral, 6 Moral–Control, and 6 Immoral–Control), presented in a semi-randomized order, balancing scenario-by-condition combinations across participants.

We also tested the possibility that patterns may differ for act-based and person-based moral judgments (Uhlmann, Zhu, & Tannenbaum, 2013). Participants were assigned (randomly, between-subjects) to make either act-based (n = 59) or person-based (n = 63) moral judgments (*Judgment Type*). For act-based judgments, participants were asked: “Are <agent>’s actions moral?” (1 – not at all, 7 – completely), and for person-based judgments, participants were asked: “Is <agent> a moral person?” (1 – not at all, 7 – completely). The name of the agent in the scenario was substituted for the placeholder text, and the same prompts were presented for first- and second-pass moral judgments.
Figure 1. Scenario design. Segments 1-4 created an initial dilemma, which was either Moral or Immoral. Segment 5 either reframed the initial dilemma, or presented morally irrelevant control information. Control information provides a baseline for changes in second-pass moral judgments. The text above is abbreviated, and 24 scenarios were used in total (see Supplemental Materials for the full text of all scenarios).

2.1.3. Procedure

Participants were instructed that they would read a series of 24 brief scenarios, that they would read each story as it unfolds, and at two points they would answer a question about it. They were told that they would be asked the same question twice, and that at each point they should answer it in light of all of the information that they currently have (see Supplemental Materials for experimental instructions). For each scenario, segments were presented cumulatively, i.e., after participants began the survey, they were presented first with segment 1, then segment 2, and so on, with each segment appearing beneath the previous segment after
participants clicked through to the next screen. After responding to all 24 scenarios, participants were asked to complete a brief demographics questionnaire.

2.1.4. Analysis

All data and analysis code are available on OSF (https://osf.io/3cyaj/?view_only=c748f6e262d947c7a5d05d709659d84a). Mixed effects analyses were performed using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2013). A linear mixed effects regression was fit to predict moral judgments, including as predictors: Initial Condition (moral, immoral), Updating Condition (reframing, control), Time Point (first-pass, second-pass), an interaction between Initial Condition and Time Point, an interaction between Updating Condition and Time Point, and the 3-way interaction. Contrasts within the model were tested using the multcomp package (Hothorn et al., 2016). Random effects parameters were chosen by first fitting a maximal model, i.e., all necessary by-subject and by-scenario random slopes and intercepts (Barr et al., 2013), then removing random effects components that showed near-zero variance in an uncorrelated model until convergence could be achieved (see Supplemental Table 1). The maximal model included random effects for Initial Condition, and Updating Condition and Time Point.

2.2. Results and Discussion

2.2.1. Act-based vs. Person-based Judgments

There were no significant differences between act-based and person-based moral judgments (see Supplemental Materials). Given that there were no significant differences between judgment types, the following analyses collapse across act-based and person-based judgments.
Figure 2. A (Study 1) & D (Study 2): Mean moral judgment for each condition, collapsed across act-based and person-based moral judgments. Error bars represent 95% confidence intervals. B (Study 1) & E (Study 2): Difference between second-pass and first-pass moral judgments, for each scenario type. C (Study 1) & F (Study 2): Magnitude of moral updating. There was greater updating for Immoral-to-Moral reframing relative to Moral-to-Immoral reframing.

2.2.2. Moral Updating

Reported p-values were corrected for multiple comparisons using the Tukey method.

Figure 2 plots mean moral ratings for each condition, and mean moral updating for each scenario type. Consistent with our intended design, Initial Moral scenarios were rated as more moral than Initial Immoral scenarios, $M = 1.30, SE = 0.13, z = 9.96, p < .001$. Also consistent with our intended design, moral judgments changed in the intended direction in both Moral-to-Immoral
and Immoral-to-Moral conditions: Moral-to-Immoral scenarios were rated as less moral than Initial Moral, $M = -0.67, SE = 0.09, z = -7.43, p < .001$, and Immoral-to-Moral scenarios were rated as more moral than Initial Immoral, $M = 1.24, SE = 0.09, z = 14.06, p < .001$. Moral judgments for Moral–Control scenarios were not significantly different from Initial Moral, $M = 0.06, SE = 0.09, z = 0.70, p = .960$. However, moral judgments for Immoral–Control were more moral than for Initial Immoral, $M = 0.24, SE = 0.09, z = 2.73, p = .037$. Critically, this positive change in moral judgment for Immoral–Control was significantly smaller than the positive change in moral judgment for Immoral-to-Moral, $M = 1.00, SE = 0.14, z = 7.40, p < .001$, meaning that moral judgments increased more from the Initial Immoral baseline when the reframing information was presented, compared to control information.

At their conclusion, Moral-to-Immoral and Immoral-to-Moral scenarios present the same exact information. Given this, do differences in the order of moral reframing lead to asymmetries in moral updating? When the absolute magnitudes of updating in Moral-to-Immoral and Immoral-to-Moral scenarios were compared, Immoral-to-Moral scenarios were updated more than Moral-to-Immoral scenarios, $M = 0.57, SE = 0.11, z = 5.34, p < .001$. That is, contrary to our prediction—based on past work documenting “negativity bias”—participants updated their moral judgments less when positive information was reframed in a negative light, and more when negative information was reframed in a positive light.

**2.2.3. Discussion**

What accounts for this unexpected asymmetry (i.e., an apparent “positivity bias” in moral updating)? One possibility is that moral and immoral scenarios are interpreted as providing different amounts of information about the situation relative to the disposition of the agent. In prior work, negativity bias has been found to stem from interpretations of negative moral
information (relative to positive information) as providing more dispositional information, i.e., immoral behaviors are infrequent, and thus more informative about a person’s character (Mende-Siedlecki, Cai, & Todorov, 2013). However, under certain conditions, information that affords strong situational attributions could still outweigh immoral dispositional information. For instance, past behavioral research has shown that, given sufficient attentional resources, situational information can override a negative dispositional impression (Fein, 1996; Fein et al., 1990; Trope & Gaunt, 2000). Study 2 directly investigated the possibility that the unexpected asymmetry in moral updating observed in Study 1 stems from differences in situational/dispositional attributions made about moral and immoral reframing information.

3. Study 2

Study 2 probed the unexpected asymmetry in moral updating by examining whether it was driven by a qualitative difference in situational/dispositional attributions about information reframing moral and immoral judgments. We hypothesized that moral reframing information may be interpreted as affording more situational, as opposed to dispositional, inferences. To this end, Study 2 replicated the design of Study 1, with an additional measure of the extent to which the scenarios allowed for situational or dispositional inferences.

3.1. Methods and Materials

3.1.1. Participants

Participants were recruited through Amazon Mechanical Turk in exchange for payment. The final sample consisted of 121 adults (56 female; $M_{\text{Age}} = 33.69$ years, $SD_{\text{Age}} = 10.07$ years), after excluding 5 participants for failing a simple attention check, and 20 participants for quitting before completing the survey. The Boston College Institutional Review Board approved the study, and each participant provided informed consent before beginning the survey.

3.1.2. Stimuli
The same 24 scenarios were used as in Study 1. For consistency with Study 1, we collected both act-based \((n = 62)\) and person-based \((n = 59)\) judgments (despite finding no differences between judgment types).

In addition to providing a moral judgment, participants were asked to make an informational judgment, addressing whether the scenario provided relatively more information about the person or about the situation. Specifically, before being prompted for a first-pass moral judgment, participants were asked: “Based on the story so far, have you learned more about <agent>, or about the situation?” (1 – Only about <agent>, 7 – Only about the situation).

Likewise, before being prompted for a second-pass moral judgment, participants were asked: “Based on the new information, have you learned more about <agent>, or about the situation?” (1 – Only about <agent>, 7 – Only about the situation; bolded emphasis in original).

3.1.3. Procedure

The procedure followed that of Study 1, with the following exception which was implemented to facilitate reading and decrease unnecessary loading delays: segments 1–4 were presented simultaneously on one page, and participants were prompted to make first-pass informational and moral judgments on this page. On the next page, segment 5 was added in bold below the previous segments, and participants were asked to make second-pass informational and moral judgments.

3.1.4. Analysis

As in Study 1, analysis was performed by fitting a linear mixed effects regression to the sample and testing contrasts within it. Random effects parameters were chosen by fitting a maximal model (Barr et al., 2013), then removing random effects components that showed near-zero variance in an uncorrelated model until convergence could be achieved (see Supplemental
Table 1). Ratings for moral and informational judgments were modeled separately. The maximal models included random effects for Initial Condition, Updating Condition, and Time Point.

Correlations between moral updating and informational judgments were also examined. To test the correlation across scenario types, a linear mixed effects model was fit to predict moral updating (difference between second-pass and first-pass moral judgments) with second-pass informational judgments as a predictor. To test the correlation within each scenario type, a linear mixed effects model was fit to predict moral updating, including as predictors: second-pass informational judgments, Initial Condition (moral, immoral), Updating Condition (reframing, control), and all interactions. The maximal model included random effects for informational judgments, Initial Condition, and Updating Condition. Lastly, a mediation analysis was conducted (*brms* package, Bürkner, 2017): a Bayesian mediation model was fit to predict moral updating among reframed scenarios, with Initial Condition (moral, immoral) as the treatment variable and second-pass informational judgments as the mediator. Default, uninformative priors were used; all Rhat values were <= 1.01, suggesting the model had converged (see Supplemental Materials for detailed model summary). The maximal model included random effects for Initial Condition and informational judgments.

### 3.2. Results and Discussion

#### 3.2.1. Act-based vs. Person-based Judgments

Replicating Study 1, there were no differences between act-based and person-based moral judgments (see Supplemental Materials). Given this, the following analyses collapse across act-based and person-based judgments.

#### 3.2.2. Moral Updating
All manipulation checks, and critically, the core Study 1 finding—that Immoral-to-Moral scenarios were updated more than Moral-to-Immoral scenarios—were replicated in the Study 2 sample (Figure 2). Reported p-values were corrected using the Tukey method. As in Study 1, all manipulation checks were passed: Initial Moral scenarios were rated as more moral than Initial Immoral, $M = 1.33, SE = 0.12, z = 11.07, p < .001$; updating occurred, such that Moral-to-Immoral scenarios were rated as less moral than Initial Moral, $M = -0.85, SE = 0.09, z = -9.51, p < .001$ and Immoral-to-Moral scenarios were rated as more moral than Initial Immoral, $M = 1.20, SE = 0.10, z = 12.46, p < .001$; control conditions showed the same pattern as in Study 1, where Moral–Control scenarios were not rated as more moral than Initial Moral, $M = 0.04, SE = 0.09, z = 0.47, p = .992$, and although Immoral–Control scenarios were rated as more moral than Initial Immoral, $M = 0.33, SE = 0.10, z = 3.38, p = .004$, this increase was significantly smaller than the corresponding increase in moral ratings for Immoral-to-Moral, $M = 0.88, SE = 0.16, z = 5.50, p < .001$. Critically, Study 2 replicated the core asymmetry observed in Study 1: the magnitude of updating was greater for Immoral-to-Moral than Moral-to-Immoral, $M = 0.35, SE = 0.12, z = 3.05, p = .014$.

### 3.2.3. Informational judgments

The core question for Study 2 was whether qualitative asymmetries in dispositional/situational information could be observed between directions of reframing (Figure 3). Contrasts compared all conditions against the scale midpoint (of four), and an additional contrast was run comparing Immoral-to-Moral and Moral-to-Immoral scenarios; p-values were corrected using the Tukey method. Initial Moral and Immoral-to-Moral scenarios were rated as providing relatively more information about the situation: Initial Moral, $M = 4.45, SE = 0.08, z = 5.92, p < .001$; Immoral-to-Moral, $M = 4.79, SE = 0.12, z = 6.58, p < .001$. Comparing these two
conditions, Immoral-to-Moral scenarios provided more situational information than Initial Moral scenarios, $M = 0.34$, $SE = 0.12$, $z = 2.90$, $p = 0.025$. None of the remaining scenarios—Initial Immoral, Moral-to-Immoral, Moral–Control, Immoral–Control—significantly differed from the scale midpoint (see Supplemental Table 2 for means). Surprisingly then, the same vignette describing a moral dilemma, whether presented initially (Initial Moral) or as reframing an immoral dilemma (Immoral-to-Moral), was judged as providing situational information. However, the vignette was seen as providing more situational information when it reframed an immoral dilemma (Figure 3). Finally, comparing Immoral-to-Moral and Moral-to-Immoral scenarios, Immoral-to-Moral scenarios were rated as providing more situational information than Moral-to-Immoral scenarios, $M = 0.93$, $SE = 0.14$, $z = 6.52$, $p < 0.001$.

![Figure 3. Study 2 condition means for informational judgments. Moral information was rated as providing more information about the situation (relative to the scale midpoint), both when it was presented in the initial segment and in the reframing segment. Moral reframing information was rated as providing more information about the situation, compared to immoral reframing information. Error bars represent 95% confidence intervals.](image)

### 3.2.4. Correlations between moral updating and informational judgments
Immoral-to-Moral scenarios elicited greater moral updating and were rated as providing more situational information compared to Moral-to-Immoral scenarios. We hypothesized that these two measures may be related, and explored the correlation between moral updating and situational (vs. dispositional) information. Moral updating was positively related to situational information, $b = 0.13$, $SE = 0.03$, $t = 4.54$, $p < 0.001$, $R^2 = 0.387$, such that, collapsing across scenario type, reframing information that was rated as providing more situational information was associated with more positive moral updating.

Similar results were found when examining within-condition correlations. Contrasts tested the relationship between moral updating and informational judgments in each condition (Figure 4); p-values were corrected using the Tukey method. Within Immoral-to-Moral scenarios, reframing information rated as providing more situational information was associated with more positive moral updating, $b = 0.11$, $SE = 0.03$, $z = 3.18$, $p = 0.006$, $R^2 = 0.297$. Likewise, within Moral-to-Immoral scenarios, reframing information rated as providing more situational information was associated with less negative moral updating, $b = 0.15$, $SE = 0.03$, $z = 4.90$, $p < 0.001$, $R^2 = 0.416$. No relationship was observed between moral updating and informational judgments among Immoral-Control scenarios, $b = 0.004$, $SE = 0.03$, $z = 0.18$, $p = 1.00$, $R^2 = 0.011$, or Moral-Control scenarios, $b = -0.007$, $SE = 0.03$, $z = -0.26$, $p = 1.00$, $R^2 = 0.002$. In sum, although moral judgment became more positive for Immoral-to-Moral scenarios and more negative for Moral-to-Immoral scenarios after reframing, the interpretation of reframing information as providing situational information biased moral updating to be more positive in both cases (Figure 4).
Figure 4. Relationship between second-pass informational judgments and moral updating for each scenario type in Study 2. For Immoral-to-Moral scenarios and Moral-to-Immoral scenarios, rating reframing information as providing more information about the situation was associated with more positive (less negative) moral updating. There was no significant relationship between informational judgments and moral updating for Immoral–Control scenarios, or for Moral–Control scenarios.

Given that (a) Immoral-to-Moral scenarios were more strongly reframed than Moral-to-Immoral scenarios, and (b) interpreting reframing information as situational was associated with more positive moral updating, we hypothesized that (b) may contribute to the "positivity bias" asymmetry observed in (a). To test this, a Bayesian mediation model was fit to predict moral updating among reframed scenarios, with Initial Condition (moral, immoral) as the treatment variable, and second-pass informational judgments as the mediator (see Supplemental Materials for detailed model summary). The mean estimated total effect of moral reframing direction on moral updating was $b = 2.08$, 95% Bayesian credible interval = [1.68, 2.50], and the mean estimated direct effect was $b = 1.97$ [1.55, 2.36]. The mean estimated indirect effect of moral reframing direction on moral updating through informational judgments was $b = 0.11$ [0.05, 0.17], representing a 5.09% [2.11%, 8.07%] mediation. These results indicate that a small
portion of the effect of moral reframing direction on moral updating is explained by informational judgments. Thus, the “positivity bias” toward stronger moral updating in the Immoral-to-Moral direction is partially explained by the interpretation of reframing information as situational.

3.2.5. Discussion

Study 2 extended and clarified the asymmetry in moral updating observed in Study 1. Despite predictions of a negativity bias, in both studies Immoral-to-Moral scenarios elicited greater updating than Moral-to-Immoral scenarios. Study 2 identified an additional asymmetry, where moral (but not immoral) reframing information was rated as providing relatively more situational (vs. dispositional) information about the agent. These asymmetries were related: for both moral and immoral reframing, situational interpretations were associated with more positive moral updating. Finally, mediation analyses indicated that the “positivity bias” toward stronger Immoral-to-Moral updating can be partially explained by moral reframing information being interpreted as providing more situational information. Together, these findings may explain why a negativity bias was overcome in both studies, and illustrate the interplay between moral judgments and qualitative interpretations of information that has the potential to reframe them.

4. General Discussion

Prior work has provided evidence for a negativity bias, where “bad is stronger than good” (Baumeister et al., 2001; Rozin & Royzman, 2001). That is, for moral judgment and impression formation, negative information is thought to be more influential than positive information. This asymmetry has been explained in terms of the relative infrequency (Mende-Siedlecki, Baron, & Todorov, 2013) and diagnosticity (Cone & Ferguson, 2015) of immoral (as opposed to moral) behaviors making immoral behaviors more important for character judgment. The present work
complicates this picture in several ways. First, Study 1 showed that the negativity bias can be reversed: moral judgments became more positive in Immoral-to-Moral scenarios than they became negative in Moral-to-Immoral scenarios, even though, after reframing, participants had access to the same information in both cases, only presented in different orders (Figure 2). Study 2 replicated this unanticipated “positivity bias” in updating and showed that it could be partly explained by an asymmetry in the qualitative interpretation of the reframing information—moral information was rated as providing more situational information, and this effect was exaggerated when moral information was presented in the context of Immoral-to-Moral (as opposed to Moral-Initial) scenarios (Figure 3). Across scenario types, rating new information as situational was associated with more positive moral updating (Figure 4), and the “positivity bias” was partially mediated by situational attributions. These findings paint a more nuanced picture of the moral judgment unfolds in the face of conflicting information.

In prior work, impression updating has been simulated by adding novel—but unrelated—anecdotes about an individual (Mende-Siedlecki et al., 2013; Brambilla et al., 2019; Gawronski et al., 2010; Petty et al., 2006). The present work used a technique that more closely resembles reality: we form impressions, then reevaluate them as we learn more (see Mann & Ferguson, 2015 for a discussion of this distinction). For example, beliefs about bad agents have been observed to actually be more volatile than beliefs about good agents (Siegel, Mathys, Rutledge & Crockett, 2018), which is consistent with our own finding that immoral first impressions can be reframed by expanding on the situational context. Critically, Siegel and colleagues collected continuous measures of character judgment over time, as one person delivered painful shocks to another in exchange for varying amounts of money. In this case, participants were given the opportunity to judge behaviors as immoral, but also to see the behavior change in the future. That
is, negative judgments can be updated if relevant counter evidence is provided. Contrast this with traditional designs, however, where counterevidence comes from completely unrelated behaviors (e.g., Laura cheated on a final exam, then later, Laura volunteered at a homeless shelter). This unrelated counterevidence may fail for a number of reasons: the original immoral action may be more threatening, it may be more diagnostic, or it may be that people don’t form generalizable impressions of someone as good or bad “on the whole”, and assessments may be more context-dependent. Future work is necessary to disentangle these hypotheses.

The present work identified an asymmetry in the qualitative interpretation of morally reframing information, where information interpreted as situational (as opposed to dispositional) facilitated more a more positive reframing of negative first impressions. This is in line with work showing that, when situational information is made salient, accessible, or specific, observers are less likely to make dispositional inferences (Fein, 1996; Fein et al., 1990; Trope & Gaunt, 2000; see Chakroff & Young, 2015 for evidence that this applies to harmful but not impure acts). In addition, in the present work moral information was operationalized as decisions in tragic dilemmas, and tragic dilemmas (c.f. taboo dilemmas), by design, involve a difficult decision (Tetlock, 2000; Mandel & Vartanian, 2008; Hanselmann & Tanner, 2008), from which an agent can still emerge having made the ‘right choice’ (Driver, 2006; Van Zyl, 2007). Interpreting tragic dilemmas as difficult situations may explain why moral reframing information elicited such high situational attributions. We speculate that moral information, used in a reframing context, can be interpreted as “pulling back the curtain”: providing necessary context that reveals a more complicated and ambiguous situation in which the initial taboo dilemma only appeared immoral because of participants’ limited access to information. That is, people may have been more amenable to positively updating their initial judgments because they had previously misjudged
the complexity of the situation. By contrast, immoral reframing information, on average, was not interpreted as providing more information about the situation, and instead only provided a conflicting signal about the character’s disposition. Note, however, that even among Moral-to-Immoral scenarios, situational ratings of immoral reframing information were associated with more positive second-pass judgments. This suggests that, even when reframing information is negative, interpreting it as revealing the broader context may contribute to a reluctance to downgrade initially positive moral judgments, at least in the context of moral dilemmas. Interestingly, even the addition of irrelevant information (in Immoral–Control scenarios) increased moral judgments relative to the Initial Immoral condition. Speculatively, in the context of moral dilemmas, even neutral information may complicate the initial impression of an agent as a purely bad actor; however, this hypothesis would require additional exploration in future work.

Our findings may have implications for how criminal defendants are perceived by juries when new information comes to light. While information about a criminal defendant’s moral behavior (e.g., volunteering at homeless shelter) may be considered immaterial to a criminal charge (Davies, 1991), jurors are expected to be sensitive to new evidence that provides additional context for the criminal act itself. Our work confirms that this recontextualizing information may correct a bad first impression, but it also suggests that jurors may be sensitive to the order in which exculpatory or incriminating evidence is presented (Shirkey, 2010). Specifically, positive information that reframes an immoral act may weigh more heavily in the final evaluations of jurors. The apparent “positivity bias” we observed may serve as an implicit acknowledgment by people that they do not have access to the complete picture. In line with this, the presumption of innocence—i.e. the acknowledgement that we cannot prematurely fill in the
gaps of our understanding—is formally enshrined in most legal systems (Tadros & Tierney, 2004), and may have an intuitive basis (Levine, Mikhail, & Leslie, 2018).

4.1. Conclusion

The present research makes a number of contributions to our understanding of how moral updating unfolds over time. First, we find that the negativity bias, well-documented by prior work in moral impression updating, does not hold in all cases, and indeed, can be reversed in some cases of ongoing moral judgment. Second, we show that this “positivity bias” appears to depend on the degree to which moral reframing information is interpreted as providing more situational context, suggesting a more complicated model of moral updating that moves beyond the simple power of valence, and instead incorporates a more holistic understanding of context and the qualitative features of informational interpretation. Critically, we identified these asymmetries by testing moral impression formation and updating in a form that more closely aligns with reality—reevaluating what we already know, rather than keeping a moral score card across independent instances. By constructing scenarios that involve moral reframing of prior behaviors, we identified important gaps in our understanding of moral impression updating, demonstrating that future research may benefit from a sensitivity to the subtle difference between the addition of new information and the reframing of what is already known.
Acknowledgments: We would like to thank Emma Alai, Kevin Jiang, Gordon Kraft-Todd, Justin Martin, Mookie Manalili, Ryan McManus, and BoKyung Park for their thoughtful feedback. This work was supported by the John Templeton Foundation (61061 to L.Y.) and the National Science Foundation (1627157 to L.Y.).

Conflicts of interest: None declared.


Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience, 33*(50), 19406-19415.


R Core Team (2013). R: A language and environment for statistical computing.


