Journal of Experimental Social Psychology 101 (2022) 104310

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Reframing of moral dilemmas reveals an unexpected "positivity bias" in updating and attributions☆

Minjae J. Kim[a,*,1], Jordan Theriault[b,1], Joshua Hirschfeld-Kroen[a], Liane Young[a]

[a] Department of Psychology and Neuroscience, Boston College, Chestnut Hill, MA, United States of America
[b] Department of Psychology, Northeastern University, Boston, MA, United States of America

ARTICLE INFO

Keywords:
Moral updating
Negativity bias
Situational/dispositional attributions

ABSTRACT

People make moral judgments about others, and new information can cause those judgments to change. Prior work examined moral impression updating after observing *additional* behaviors, but less is known about moral updating when prior behaviors are *reframed*, either as more or less moral than on first impression. The present work compared moral updating as scenarios were reframed from moral-to-immoral, or immoral-to-moral. Three studies show that the negativity bias, well-documented by prior work, can be reversed when first impressions are reframed, and partially reinstated if new information is irrelevant. Further, this "positivity bias" is partly explained by the extent to which reframing information elicits external causal attributions. Future research on moral updating may benefit from a sensitivity to such qualitative features of new information.

1. Introduction

Moral judgment is a dynamic process: we form initial impressions of others, and we update these impressions as new information comes to light. A large body of social psychological research has identified factors that influence moral impression updating, where initial impressions are revised in the face of counter-attitudinal information (e.g., Brambilla, Carraro, Castelli, & Sacchi, 2019; Klein & O'Brien, 2016; Mende-Siedlecki, Baron, & Todorov, 2013; Monroe & Malle, 2019; Reeder & Coovert, 1986). Typically, this work presents participants with sequences of disparate behaviors that change in valence over time (e.g., in sequence: "Laura translated items for a foreigner in a restaurant", "Laura gave money to charity", "Laura heckled a woman speaking on human rights"; Mende-Siedlecki, Baron, & Todorov, 2013). Such paradigms, where new information about a target is *unrelated to* old information, may be called "addition" paradigms (Mann & Ferguson, 2015).

In our day-to-day experience, however, revisions of moral judgments often occur in response to relevant information: we may come across new details about someone's good or bad behavior that prompt us to revisit our initial inference based on that behavior. A growing body of work has examined how moral impression updating unfolds when new information about a target is *related to* old information. Such paradigms may be broadly called "reframing" paradigms. For instance, Mann and Ferguson (2015) have shown that implicit and explicit negative impressions can be successfully undone if new information affords a positive *reinterpretation* of the target's previous behavior. To our knowledge, however, no empirical work has systematically compared two possible directions of reframing: positive reframing of an initially negative impression, versus negative reframing of an initially positive impression. Such work could identify an asymmetry, where some initial impressions are harder to revise than others. Prior work using addition paradigms has revealed a negativity bias in impression updating, where "bad is stronger than good" (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Cone & Ferguson, 2015; Mende-Siedlecki, Baron, & Todorov, 2013; Rozin & Royzman, 2001; Skowronski & Carlston, 1989). The present work examines whether a negativity bias persists in a *reframing* paradigm: will moral updating be stronger when an initially positive impression gets reframed as negative, or when an initially negative impression gets reframed as positive? We incorporate insights from attribution theory (Heider, 1958; Reeder & Brewer, 1979; Gawronski & Brannon, 2019) into moral updating, providing a potential explanation for both currently observed results and previous results, including the disproportionate influence of negative information on updating.

☆ This paper has been recommended for acceptance by Rachel Barkan.

* Corresponding author.
E-mail address: minjae.kim@bc.edu (M.J. Kim).

[1] Equal contributors.

https://doi.org/10.1016/j.jesp.2022.104310
Received 12 July 2021; Received in revised form 22 February 2022; Accepted 28 February 2022
Available online 21 March 2022

1.1. It is easier to make someone look worse, than to make someone look better

Prior work using addition paradigms has established the presence of a negativity bias in impression formation, such that negative information is more influential than positive information (Baumeister et al., 2001; Reeder & Coovert, 1986; Rozin & Royzman, 2001; Skowronski & Carlston, 1987). For example, when observers are provided with both positive and negative trait information about a target, the unfavorable traits contribute more to global impression ratings than a simple averaging model would predict (Anderson, 1965). The negativity bias has also been demonstrated in the more specific context of moral updating: when forming an impression of a target's moral character (e.g., *how moral* or *how trustworthy*) from separate behavioral examples, participants revise their impressions more when moral examples are followed by immoral examples, compared to the opposite (Kim, Mende-Siedlecki, Anzellotti, & Young, 2021; Mende-Siedlecki, Baron, & Todorov, 2013; Reeder & Coovert, 1986).

Furthermore, prior work using addition paradigms has identified several features that predict the negativity bias. For one, perceived behavioral frequency matters: compared to moral and incompetent behaviors, immoral and competent behaviors are perceived to occur less frequently, and elicit stronger updating of explicit impressions (Mende-Siedlecki, Baron, & Todorov, 2013). These results support a cue-diagnosticity mechanism (Skowronski & Carlston, 1987), where the negativity bias stems from the relative rarity of immorality (rather than negativity per se). Likewise, judgments of diagnosticity contribute to a negativity bias in implicit character evaluations (evaluations that are automatically activated and indirectly measured; Greenwald & Banaji, 1995; Greenwald, McGhee, & Schwartz, 1998), such that an extremely negative behavior (mutilating a small animal) moves initially positive evaluations more than an extremely positive behavior (donating a kidney) moves initially negative evaluations (Cone & Ferguson, 2015). Importantly, participants who deemed the negative behavior as more offensive also judged the behavior to be more reflective of true character, and this diagnosticity judgment predicted the degree of implicit impression updating (Cone & Ferguson, 2015).

1.2. Negative impressions can be undone through reframing

Of course, despite the negativity bias, we sometimes do judge people more positively after learning new information. Prior work suggests that we are likelier to do this when the new information can reframe what was learned initially. In a series of studies, Mann and Ferguson (2015) compared changes in implicit evaluations between a reframing condition (where new information was related to previous behavior) and an addition condition (where new information was unrelated to previous behavior). They found that negative implicit evaluations based on seemingly bad behavior can be successfully reversed if new information can be used to *reinterpret* the previous behavior as positive, but not if new information takes the form of an additional, unrelated positive behavior. Specifically, participants first read about a man who violently broke into his neighbors' homes and took precious things from them; then, participants in the reframing condition read new information that revealed that the neighbors' homes were on fire, and that the "precious things" he removed were young children, while participants in the addition condition read an unrelated but equally positive story about the same man (where he saved a baby who had fallen onto the train tracks). Participants who read about the fire rescue exhibited a successful update from a negative evaluation to a positive one; those who read the unrelated story only exhibited an attenuated negative evaluation. These results suggest that, at least for implicit evaluations, impression change can occur flexibly in the positive direction, but requires that new information be able to change the meaning of what was previously shown. This is critical because, as reviewed above, prior work revealing a negativity bias has used addition paradigms to probe updating, where disparate behaviors are presented in sequence.

1.3. Modes of reframing

When new information about a target is meaningfully related to old information about the same target, this relatedness may take different forms. As just described, new information may prompt *reinterpretation*—a change in meaning—of old information (Mann & Ferguson, 2015). Another potential mechanism (that may co-occur with the first) is *recontextualization*—a broadening of context. For instance, after we learn that someone has done something wrong, we may learn a new motive for the behavior that may provide a reasonable (or at least partial) justification for that behavior. In this case, we may shift our attribution of that person's behavior from more dispositional (arising from their stable traits) to more situational (arising from the situation; Heider, 1958); this change in attribution may in turn shift our moral judgment.

Reinterpretation and recontextualization both require a meaningful relationship between new and old information, such that the new information prompts us to revisit our initial inference in some way. In the present work, the experimental paradigm focuses on recontextualization, but it makes no direct comparisons between recontextualization and reinterpretation (which need not be mutually exclusive; we only note the two for ease of characterizing our paradigm in relation to prior work). For the sake of simplicity throughout this paper, we will refer to all paradigms where related pieces of information are presented as "reframing" paradigms—in contrast to "addition" paradigms—regardless of the exact process that may be captured by the paradigm. Furthermore, we recognize that, in addition paradigms, new and old information tend to be opposite in valence (clearly positive and negative), while in reframing paradigms, new and old information may not be best described as being "opposite in valence", so terms like "negativity bias" and "moral updating" may have a better conceptual fit to addition paradigms. That said, in the present work, we are still interested in comparing different directions of reframing; thus, we will refer to these directions using valenced language, and we will use the term "moral updating" to refer to *changes in moral judgment* that are measured both in addition paradigms and in reframing paradigms.

1.4. Present work

The present work uses a set of brief narratives to explore explicit moral updating in a reframing paradigm, where new information reveals a counter-valenced motive underlying the agent's previous behavior. Two directions of reframing are compared: positive-to-negative and negative-to-positive. Critically, the scenarios consist of interchangeable segments that can be arranged to produce both reframing directions; this means that the exact same information gets presented in both conditions, only in different orders. Does a negativity bias in moral updating still operate in this reframing paradigm, as in prior work using addition paradigms? Or will moral updating proceed along a different path?

All scenarios in the present work first framed someone's behavior as relatively moral or relatively immoral, and later revealed a new motive for the behavior that was relatively immoral or relatively moral. Scenarios drew on prior examples of tragic and taboo dilemmas (Fiske & Tetlock, 1997; Tetlock, Kristel, Elson, Green, & Lerner, 2000; Tetlock, 2002, 2003; McGraw & Tetlock, 2005; Ginges, Atran, Medin, & Shikaki, 2007; Lichtenstein, Gregory, & Irwin, 2007; Bartels, 2008; Hanselmann & Tanner, 2008; Mandel & Vartanian, 2008; Tannenbaum, Uhlmann, & Diermeier, 2011). 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 choose between two competing moral values, such that they must choose the "lesser of two evils" (Fig. 1c "Moral"). Imagine a scenario used in the present work: a fishing boat captain must decide whether or not to buy expensive new fishing

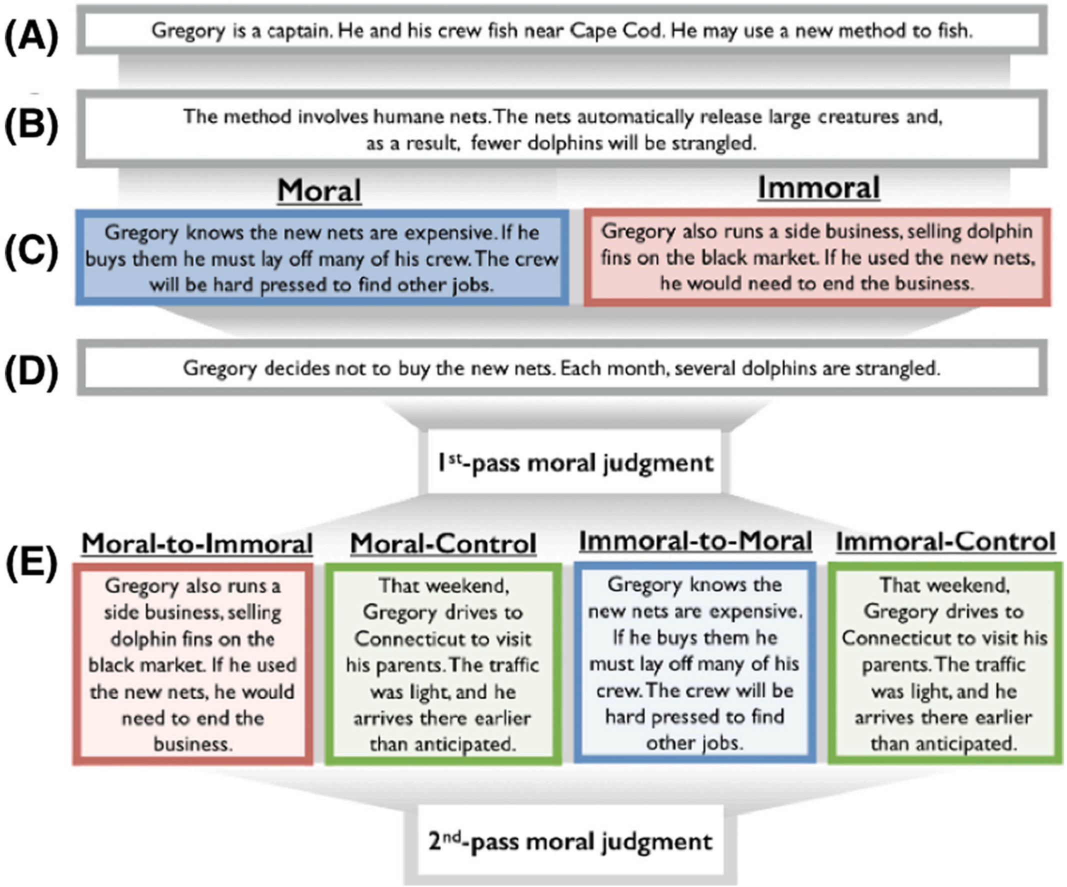
(A)
Gregory is a captain. He and his crew fish near Cape Cod. He may use a new method to fish.
(B)
The method involves humane nets. The nets automatically release large creatures and, as a result, fewer dolphins will be strangled.
Moral
Immoral
(C)
Gregory knows the new nets are expensive. If he buys them he must lay off many of his crew. The crew will be hard pressed to find other jobs.
Gregory also runs a side business, selling dolphin fins on the black market. If he used the new nets, he would need to end the business.
(D)
Gregory decides not to buy the new nets. Each month, several dolphins are strangled.
1st-pass moral judgment
(E)
Moral-to-Immoral
Moral-Control
Immoral-to-Moral
Immoral-Control
Gregory also runs a side business, selling dolphin fins on the black market. If he used the new nets, he would need to end the business.
That weekend, Gregory drives to Connecticut to visit his parents. The traffic was light, and he arrives there earlier than anticipated.
Gregory knows the new nets are expensive. If he buys them he must lay off many of his crew. The crew will be hard pressed to find other jobs.
That weekend, Gregory drives to Connecticut to visit his parents. The traffic was light, and he arrives there earlier than anticipated.
2nd-pass moral judgment

Fig. 1. Scenario design. Parts *a–d* created an initial dilemma, which was either moral or immoral. Part *e* either reframed the initial dilemma, or presented morally irrelevant control information. Control information provides a baseline for changes in moral judgments. The text above is abbreviated, and 24 scenarios were used in total (see Supplemental Materials for the full text of all scenarios).

nets that would kill fewer dolphins, but to compensate for the expense, he would also need to lay off many of his crew (who would not find jobs easily). The ambiguity of tragic dilemmas makes moral judgment difficult, as it is unclear which choice is "morally correct" (Tetlock et al., 2000); as such, if the captain decides to not buy the new nets, thus protecting the livelihoods of his workers, he may be judged as relatively more moral. By comparison, in *taboo dilemmas*, decision-makers choose between a moral concern and economic self-interest (Fig. 1c "Immoral"). Imagine a fishing boat captain who runs a side-business selling dolphin fins on the black market; when he learns about the new nets, he must decide between keeping his side-business and saving dolphins. Most people believe that prioritizing economic self-interest over the lives of others is wrong (Tetlock, 2003); as such, if the captain decides to not buy the new nets, thus maximizing his profit at the expense of the dolphins, he may be judged as relatively more immoral. Critically, tragic and taboo cases can be reframed (Atran & Axelrod, 2008): what first looked like a relatively moral decision in the context of a tragic dilemma can be reframed by revealing a hidden economic incentive behind the "less evil" option (Fig. 1e "Moral-to-Immoral"), and what first looked like a relatively immoral decision in the context of a taboo dilemma can be reframed by revealing a secondary moral good produced by the "selfish" option (Fig. 1e "Immoral-to-Moral").

All scenarios in the present work were composed of interchangeable segments (Fig. 1) that can be arranged to proceed along two paths of experimental interest: *Moral-to-Immoral* or *Immoral-to-Moral*. In the tragic and taboo dilemmas that these scenarios reflect, a tragic dilemma involves choosing between two outcomes that both have moral value, and a taboo dilemma involves one outcome with moral value and one without. As a consequence, agents' "moral" decisions in a tragic dilemma are better described as *relatively more moral*, in contrast to the less ambiguously "immoral" decisions in a taboo dilemma, since a tragic dilemma forces choice between two arguably "moral" outcomes. However, for linguistic simplicity, we use the terms "moral" and "immoral" for all condition labels, and throughout this paper. For every story presented in the experiment, participants provided two moral judgments: a first-pass moral judgment after the initial framing, and a second-pass moral judgment after the reframing. We examined moral updating between the first-pass judgment and the second-pass judgment.

In all, three experiments were conducted to compare positive-to-negative and negative-to-positive moral updating in a reframing paradigm, and test a potential mechanism for the observed asymmetry. In Study 1, contrary to prior work using addition paradigms, we observed greater updating for Immoral-to-Moral scenarios than Moral-to-Immoral scenarios. We hypothesized that this positivity bias may be related to differences in dispositional and situational attributions between moral and immoral information (Fein, 1996; Fein, Hilton, & Miller, 1990; Gawronski, 2004; Gilbert & Jones, 1986; Gilbert & Malone, 1995). Study 2 investigated this asymmetry, revealing that new moral information was perceived as providing more situational information than new immoral information; further, situational interpretations predicted stronger positive updating. Next, Study 3 compared our reframing paradigm to an addition paradigm, showing that the positivity bias goes away when new information is unrelated to old information. Study 3 also dissected the relationship between situational interpretations and moral updating, and showed that perceptions of a behavior as being *externally* caused are particularly relevant for updating. These studies indicate how "zooming out" to provide a new, external motive can recontextualize moral judgments, making what at first appeared relatively immoral seem more moral in the end.

2. Study 1

Study 1 examined asymmetries in moral updating between Moral-to-Immoral and Immoral-to-Moral scenarios in a reframing paradigm. This experiment was not preregistered, but all key effects replicate in Studies 2–3, which were preregistered.

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 (59 identified as male, 63 identified as 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. After collecting data from all 122 participants, we conducted our analyses without collecting additional data. We did not conduct a formal power analysis to determine our sample size, but we took into consideration the stimuli and subjects necessary to detect effects in mixed effects designs, according to prior simulations (Westfall, Kenny, & Judd, 2014). Our sample size is powered to detect a minimum effect size of $d = 0.256$ at $\alpha = 0.05$, $\beta = 0.80$, in a paired sample *t*-test (Faul, Erdfelder, Lang, & Buchner, 2007). Please see footnote 1 for a more specific sensitivity analys tailored to our mixed effects design. The study was approved by the Boston College IRB, and each participant provided informed consent before beginning the survey. While analyses were conducted with a subset of measures and participants, all measures, manipulations, and exclusions are reported.

2.1.2. Stimuli

Stimuli consisted of 24 detailed moral scenarios, either adapted from prior work (Critcher, Helzer, Tannenbaum, & Pizarro, 2012; Tetlock et al., 2000; Uhlmann, Zhu, & Tannenbaum, 2013) or created for the current study. Each scenario began by introducing an agent, and presenting their tragic (relatively *moral*) or taboo (relatively *immoral*) dilemma (*Initial Condition*). Following this, each scenario presented a final piece of information (*Reframing Condition*), designed to either reframe the initial information (*reframing*), or to be morally irrelevant (*control*) (Fig. 1). Participants read each scenario and made moral judgments at two time points (*Timepoint*): a *first-pass* judgment after presentation of the *Initial Condition* segment, and a *second-pass* judgment after presentation of the *Reframing Condition* segment.

The combination of Initial Condition, Reframing Condition, and Timepoint created six conditions: (1) Initial Moral, (2) Initial Immoral, (3) Moral-to-Immoral, (4) Immoral-to-Moral, (5) Moral–Control, (6) Immoral–Control. Each root scenario was composed of pieces that could be rearranged to generate all six conditions. This design also ensured that second-pass judgments of Moral-to-Immoral and Immoral-to-Moral scenarios were made using *the exact same information*, *only presented in different orders* (Fig. 1). This design is important, as prior work has shown that extremity can drive moral judgments (Cone, Mann, & Ferguson, 2017; Reeder & Coovert, 1986), but in the present case, the information available is controlled. 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 to balance scenario-by-condition combinations between participants.

We also tested whether moral updating was differently affected for act-based and person-based moral judgments (Uhlmann et al., 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 same prompts were presented for first- and second-pass moral judgments.

2.1.3. Procedure

Participants were instructed that they would read 24 brief stories as they unfold, and that at two points in each story 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, each part (*a* through *d*) appeared beneath the previous part once participants clicked through to the next screen. After responding to all 24 scenarios, participants completed a brief demographics questionnaire.

2.1.4. Analysis

All data and analysis code are available on OSF (see Open Practices). Mixed effects analyses were performed using the *lme4* package in R (Bates, Mächler, Bolker, & Walker, 2014; R Core Team, 2013). A linear mixed effects regression was fit to predict moral judgments, using as predictors: Initial Condition (moral, immoral), Reframing Condition (reframing, control), Timepoint (first-pass, second-pass), and all interactions except Initial Condition x Reframing Condition, as this interaction is excluded by the design (as there is no difference between reframing and control conditions during the first-pass judgment). *P*-values for fixed effects were obtained via Satterthwaite’s degrees of freedom method in the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017). We report partial Eta-squared values as effect sizes for interaction terms (obtained using the *effectsize* package; Ben-Schachar, Lüdecke, & Makowski, 2020). Contrasts within the model were tested simultaneously using the *multcomp* package (Hothorn et al., 2016). Effect sizes (Cohen’s *d*) for contrasts were estimated by dividing the mean difference by the square root of the summed variance components (Brysbaert & Stevens, 2018). All reported *p*-values are corrected for multiple comparisons using the Tukey method.

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, Levy, Scheepers, & Tily, 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, Reframing Condition, and Timepoint.

2.2. Results

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.

2.2.2. First-pass moral judgments

Consistent with our design, Initial Moral segments ($M = 4.01$, $SE = 0.18$) were rated as more moral than Initial Immoral segments ($M = 2.71$, $SE = 0.17$), *Estimate* = 1.30, $SE = 0.13$, $z = 9.96$, $p < 0.001$, $d = 0.54$ [95% Confidence Interval = 0.43, 0.65]. Initial Immoral segments received more extreme moral ratings than Initial Moral segments: Initial Moral ratings ($M = 4.01$, $SE = 0.18$) did not significantly differ from the scale midpoint of 4, $z = 0.03$, $p = 1.00$, $d = 0.002$ [−0.145, 0.150], whereas Initial Immoral ratings ($M = 2.71$, $SE = 0.17$) were significantly lower than the midpoint, $z = 7.76$, $p < 0.001$, $d = 0.53$ [0.40, 0.67] (Fig. 2a).

2.2.3. Moral updating

Reframing in both directions was successful, consistent with our design. Contrast analyses revealed that Initial Moral segments ($M = 4.01$, $SE = 0.18$) became less moral when reframed to Moral-to-Immoral

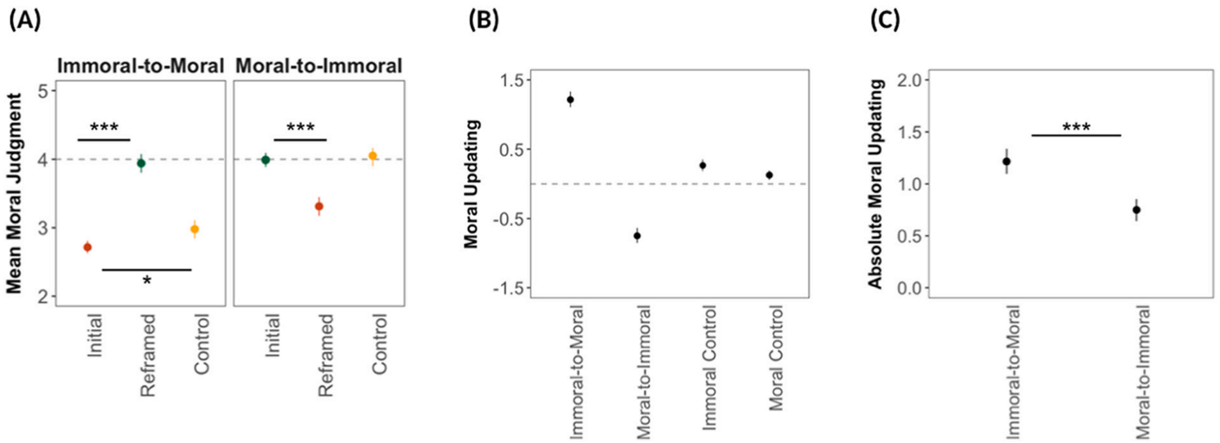

Fig. 2. Study 1 mean moral judgments and updating. A: Mean moral judgment for each condition, collapsed across act-based and person-based moral judgments. Error bars represent 95% confidence intervals. B: Difference between second-pass and first-pass moral judgments, for each scenario type. C: Magnitude of moral updating. There was greater updating for Immoral-to-Moral reframing relative to Moral-to-Immoral reframing. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ (after correction for multiple comparisons).

($M = 3.34$, $SE = 0.18$), *Estimate* $= -0.67$, $SE = 0.09$, $z = -7.43$, $p < 0.001$, $d = -0.28$ [−0.35, −0.21]. Initial Immoral segments ($M = 2.71$, $SE = 0.17$) became more moral when reframed to Immoral-to-Moral ($M = 3.95$, $SE = 0.17$), *Estimate* $= 1.24$, $SE = 0.09$, $z = 14.06$, $p < 0.001$, $d = 0.52$ [0.44, 0.59]. Moral judgments of Initial Moral segments remained unchanged after adding Control information ($M = 4.07$, $SE = 0.21$), *Estimate* $= 0.06$, $SE = 0.09$, $z = 0.70$, $p = 0.960$, $d = 0.03$ [−0.05, 0.10]; however, Initial Immoral segments did become slightly more moral after adding Control information ($M = 2.95$, $SE = 0.19$), *Estimate* $= 0.24$, $SE = 0.09$, $z = 2.73$, $p = 0.037$, $d = 0.10$ [0.03, 0.17]. This moral shift for Immoral–Control scenarios was significantly smaller than the equivalent shift for Immoral-to-Moral scenarios, *Estimate* $= 1.00$, $SE = 0.14$, $z = 7.40$, $p < 0.00$, $d = 0.42$ [0.31, 0.53].

At their end, Moral-to-Immoral and Immoral-to-Moral scenarios both present participants with the same exact information, only presented in a different order. We asked whether differences in order created asymmetries in moral updating. There was a significant 3-way interaction[2] between Initial Condition, Reframing Condition, and Timepoint, *Estimate* $= 1.73$, $SE = 0.23$, $t(27.52) = 7.38$, $p < 0.001$, $\eta^2_p = 0.66$ [0.43, 0.79]. We observed a positivity bias when comparing the absolute magnitudes of moral updating in Moral-to-Immoral ($M = 3.34$, $SE = 0.18$) and Immoral-to-Moral ($M = 3.95$, $SE = 0.17$) scenarios: Immoral-to-Moral scenarios were updated *more* than Moral-to-Immoral scenarios, *Estimate* $= 0.57$, $SE = 0.11$, $z = 5.34$, $p < 0.001$, $d = 0.24$ [0.15, 0.33].

2.3. Discussion

Contrary to prior work using addition paradigms, in our reframing paradigm, we observed a positivity bias, where there was greater updating for Immoral-to-Moral scenarios compared to Moral-to-Immoral scenarios. That is, participants updated their moral judgments *more* when a given a justification for what appeared selfish, and updated *less* when made aware of a selfish ulterior motive. What could account for this unexpected asymmetry (an apparent positivity bias in moral updating)?

2.3.1. Do moral and immoral information show differences in extremity?

Prior work has shown that extremity can drive moral judgments and updating (Cone et al., 2017; Reeder & Coovert, 1986); however, an important feature of the current design is that Moral-to-Immoral and Immoral-to-Moral scenarios presented the exact same information, in different orders. Thus, in theory, the extremity of *all available information* was controlled between these scenario types. That said, it is possible that moral reframing information, on average, was perceived to be more extreme than immoral reframing information. This was not the case, however: we tested the extremity of moral ratings for Initial Moral and Initial Immoral segments (***First-pass moral judgments***), and found that Initial Moral segments were judged as *less extreme* (closer to the midpoint) than Initial Immoral segments. This is in opposition to the direction of the observed effect, and thus it is unlikely that perceived extremity induced greater moral reframing in the present design. Furthermore, it is noteworthy that judgments of Initial Moral segments did not significantly differ from the scale midpoint—this highlights the more ambiguous nature of tragic dilemmas, and the difficulty of adjudicating between two arguably moral options (Tetlock et al., 2000).

2.3.2. Do moral and immoral information show differences in frequency?

Another candidate explanation for the observed positivity bias is differences in perceived behavioral frequency (Cialdini, Reno, & Kallgren, 1990; Fiske, 1980; Mende-Siedlecki, Baron, & Todorov, 2013; Reeder, 1993). It is possible that moral reframing information, on average, was perceived to be less frequent than immoral reframing information, and that this difference contributed to greater updating.

To rule out this possibility, a new group of participants was recruited to provide first- and second-pass ratings for all 24 scenarios along five features, including descriptive frequency ($N = 62$; see Supplementary Materials for methods and full list of features; data available on OSF, see Open Practices). In second-pass ratings for Moral-to-Immoral and Immoral-to-Moral scenarios, we observed no difference in descriptive frequency (Moral-to-Immoral: $M = 4.12$, $SE = 0.02$; Immoral-to-Moral: $M = 4.17$, $SE = 0.02$; *Estimate* $= 0.04$, $SE = 0.03$, $t(1440) = 1.57$, $p = 0.117$, $d = 0.04$ [−0.01, 0.09]). Given this, it is unlikely that moral reframing behaviors induced greater updating because they were perceived to be less frequent. A more probable explanation for the valence asymmetry in updating may lie elsewhere.

[2] A sensitivity analysis (estimated by simulation using the *simr* package; Green & MacLeod, 2016) indicated that the 3-way interaction between Initial Condition, Reframing Condition, and Timepoint could be detected at a minimum effect size 60% below the observed effect size, while retaining ~80% power (Arend & Schäfer, 2019; Bloom, 1995). All fixed effects in the model were multiplied by 0.4, and a Monte Carlo simulation was used to conduct a z-test on the interaction term (*power* = 83.30% [80.84%, 85.56%], 1000 simulations, function call: powerSim(model, nsim = 1000, test = fixed(“Initial:Updating:Timepoint”, method = “z”), seed = 123)).

2.3.3. Situational and dispositional attributions may explain the positivity bias

One potential explanation for the positivity bias is that moral and immoral reframing information are interpreted as differently providing dispositional and situational information. Reeder and Brewer's (1979) theory of attribution suggests that immoral behaviors are typically attributed to the agent's disposition, and moral behaviors tend to get attributed to the situation. They hypothesize that this attribution pattern stems from a hierarchical schema, where moral behaviors are regularly performed by both moral and immoral people, but only immoral people perform immoral behaviors. Extending this schema to the context of moral updating may explain the negativity bias observed in prior work using addition paradigms: immoral behaviors that violate moral impressions lead to more dispositional attributions and greater belief updating, while moral behaviors that violate immoral impressions lead to more situational attributions and reduced belief updating (Gawronski & Brannon, 2019).

How might situational and dispositional attributions also explain the positivity bias observed in the current paradigm? Situational attributions may lead to different outcomes for updating, depending on the relationship between new and old information. If the new, impression-violating information is *unrelated* to past behaviors, then the agent's new behavior may be attributed to the situation—i.e., perceived as arising from a non-dispositional cause. For example, if a person who was previously described as untrustworthy donates money to charity, we may assume they did it to look good, not because their disposition changed. On the other hand, if the new, impression-violating information is directly related to a past behavior, the new information may be perceived as revealing the situational constraints on the agent's past behavior. For example, when it is revealed that Gregory the sea captain might have had to lay off his crew (Fig. 1e), we may be more understanding of his decision to use cheap but dolphin-harming nets. These contrasting attribution processes—attributing a new behavior to the situation vs. inferring situational constraints on an old behavior—can both be paths through which new information can be reconciled with an existing impression (Brannon & Gawronski, 2018). Situational attributions, then, may explain not just the negativity bias previously observed in addition paradigms, but also the positivity bias observed in the current reframing paradigm.

We hypothesized that moral reframing information, compared to immoral reframing information, will be perceived as providing more information about the situation (surrounding the agent's decision), and that this will in turn predict moral updating. Study 2 directly investigated this hypothesis.

3. Study 2

Study 2 was a preregistered (https://osf.io/8j63q) experiment that replicated the design of Study 1, and extended it by including a measure of the extent to which participants learned about the agent's disposition vs. the situation. Study 2 was preceded by an initial, exploratory follow-up experiment to Study 1 with the exact methods that will be described below; however, as neither Study 1 nor the initial follow-up were preregistered, we ran a preregistered direct replication of both experiments, reported here as Study 2 (results from the original follow-up experiment are reported in Supplemental Materials as Study 1.5).

Our main hypotheses were as follows (based on findings from Study 1 and the initial follow-up): (1) the magnitude of moral updating will be greater for Immoral-to-Moral vs. Moral-to-Immoral scenarios; (2) moral reframing information will be rated as providing more situational information than immoral reframing information; (3) reframing information that is rated as more situational will be associated with more positive updating; and (4) situational interpretations will partially mediate the positivity bias in moral updating.

3.1. Methods and materials

3.1.1. Participants

The sample size was preregistered to include 550 participants (275 per moral judgment type), and participants were excluded if they quit before completing the survey, failed an attention check ("When the following numbers are arranged by their numerical value, which is the middle number: two, eight, or three?"), or reported post-task that they did not attend to the task (responding <4 on a 1–7 scale, 1 = paid almost no attention, 7 = paid all my attention). Participants were recruited through Amazon Mechanical Turk in exchange for payment. The final sample consisted of 549 adults (265 identified as male, 278 identified as female, 3 identified as non-binary/other; $M_{\text{Age}} = 41.16$ years, $SD_{\text{Age}} = 11.92$ years), after excluding 27 participants for failing the attention check, 66 participants for reporting that they did not attend to the task, and 21 participants for quitting before completing the survey. After collecting data from all 549 participants, we conducted our analyses without collecting additional data. For sensitivity analyses tailored to the mixed effects design for this study, see footnotes 2, 4, and 5. The study was approved by the Boston College IRB, and each participant provided informed consent before beginning the survey. While analyses were conducted with a subset of measures and participants, all measures, manipulations, and exclusions are reported.

3.1.2. Stimuli

The same 24 scenarios were used as in Study 1. As in Study 1, act-based ($N = 276$) and person-based ($N = 273$) moral judgments were collected.

Participants were asked to make a moral judgment as well as an *informational judgment* about whether the scenario provided relatively more information about the person or about the situation. Just before their 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, just before their 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). This question probed the information gained, rather than attribution made (e.g., "Is this behavior due to the agent or the situation?"), as the reframing information elaborated on the initial behavior, rather than presenting an additional, new behavior where attribution could be probed.

3.1.3. Procedure

The procedure followed that of Study 1, with one change to increase usability for participants: parts *a* through *d* (Fig. 1) were presented on one page, where participants made first-pass informational and moral judgments. On the next page, part *e* was added in bold below the previous parts, and participants made second-pass informational and moral judgments.

3.1.4. Analysis

All data and analysis code are available on OSF (see Open Practices). Linear mixed effects regressions were separately fit to predict moral judgments and informational judgments. Predictors included Initial Condition (moral, immoral), Reframing Condition (reframing, control), Timepoint (first-pass, second-pass), and all interactions except Initial Condition x Reframing Condition (as in Study 1), and the random effects structure was reduced according to the procedure outlined in Study 1 (see Supplemental Table 1). The maximal models included random effects for Initial Condition, Reframing Condition, and Timepoint. All reported *p*-values are corrected for multiple comparisons using the Tukey method.

A separate linear mixed effects model tested correlations between moral updating and informational judgments. Moral updating (second-pass minus first-pass moral judgments) was predicted using, as predictors, second-pass informational judgments, Initial Condition (moral, immoral), Reframing Condition (reframing, control), and all interactions. The maximal model included random effects for informational judgments, Initial Condition, and Reframing Condition. All reported p-values are corrected for multiple comparisons using the Tukey method.

Finally, second-pass informational judgments were tested as a mediator between Initial Condition (moral, immoral) and moral updating in a Bayesian multilevel model (*brms* package; Bürkner, 2017). Default, uninformative priors were used, and all Rhat values were $< =$ 1.01, suggesting the model had converged. The maximal model included random effects for Initial Condition and informational judgments.

3.2. Results

3.2.1. Act-based vs. person-based judgments

We compared act-based and person-based judgments for all conditions. For Immoral–Control segments, act-based moral judgments ($M = 2.99$, $SE = 0.20$) and person-based moral judgments ($M = 2.75$, $SE = 0.20$) were significantly different ($Estimate = 0.24$, $SE = 0.09$, $z = 2.80$, $p = 0.027$, $d = 0.10$ [0.03, 0.18]). There were no other significant differences between act-based and person-based moral judgments (see Supplemental Materials). As Immoral–Control scenarios are not involved in our main hypotheses (which all concern reframed scenarios), we collapsed across act-based and person-based judgments for the following analyses (all qualitative patterns of results replicate within each judgment type; see Supplemental Materials).

3.2.2. Moral updating

Consistent with Study 1, a valence asymmetry was present in the magnitude of moral updating, such that immoral-to-moral reframing showed a larger absolute change in judgments, compared to moral-to-immoral reframing. A significant 3-way interaction[3] was observed between Initial Condition, Reframing Condition, and Timepoint, $Estimate = 2.13$, $SE = 0.18$, $t(25.10) = 11.72$, $p < 0.001$, $\eta^2_p = 0.85$ [0.71, 0.90], and the absolute magnitude of updating was greater for Immoral-to-Moral ($M = 1.38$, $SE = 0.05$) than Moral-to-Immoral ($M = 0.97$, $SE = 0.05$) scenarios, $Estimate = 0.41$, $SE = 0.05$, $z = 8.79$, $p < 0.001$, $d = 0.18$ [0.14, 0.22] (Fig. 3c).

Mean moral ratings for each segment (Fig. 3a) were all consistent with the experimental manipulation (see Supplemental Materials for further comparisons and statistics): (1) Initial Moral segments were rated as more moral than Initial Immoral; (2) Moral-to-Immoral segments were rated as less moral than Initial Moral; (3) Immoral-to-Moral segments were rated as more moral than Initial Immoral; and (4) while Immoral-Control segments were rated as more moral than Initial Immoral, this update was smaller than the update in Immoral-to-Moral scenarios.[4]

3.2.3. Informational judgments

An asymmetry was present in informational judgments that corresponded to the asymmetry in reframing: Immoral-to-Moral segments ($M = 4.80$, $SE = 0.13$) provided more situational information than Moral-to-Immoral segments ($M = 3.75$, $SE = 0.10$), $Estimate = 1.05$, $SE = 0.16$, $z = 6.45$, $p < 0.001$, $d = 0.38$ [0.26, 0.49] (Fig. 4). This comparison was observed after following up a significant 3-way interaction[5] between Initial Condition, Reframing Condition, and Timepoint, $Estimate = 1.19$, $SE = 0.17$, $t(23.85) = 6.86$, $p < 0.001$, $\eta^2_p = 0.66$ [0.41, 0.79] (see Supplemental Materials for comparisons of informational judgments against the scale midpoint.)

3.2.4. Correlations between moral updating and informational judgments

Moral updating and second-pass informational judgments were correlated (Fig. 5), such that within Immoral-to-Moral scenarios, reframing information that provided *more situational* information was associated with *more positive* moral updating, $Estimate = 0.09$, $SE = 0.02$, $z = 5.79$, $p < 0.001$, $d = 0.05$ [0.03, 0.07], and within Moral-to-Immoral scenarios, reframing information that provided *more situational* information was associated with *less negative* moral updating, $Estimate = 0.16$, $SE = 0.01$, $z = 11.40$, $p < 0.001$, $d = 0.09$ [0.07, 0.10]. No correlation was observed in Immoral-Control scenarios, $Estimate = -0.017$, $SE = 0.011$, $z = -1.615$, $p = 0.435$, $d = -0.0002$ [−0.0012, 0.0114], or in Moral-Control scenarios, $Estimate = -0.0003$, $SE = 0.0106$, $z = -0.027$, $p = 0.999$, $d = -0.0002$ [−0.0117, 0.0114]. Thus, when reframing information was interpreted as situational, moral judgments became more positive. These comparisons were observed after following up a significant 2-way interaction[6] between informational judgments and Reframing Condition, $Estimate = 0.13$, $SE = 0.01$, $t(12710) = 10.39$, $p < 0.001$, $\eta^2_p = 0.008$ [0.006, 0.012] (see Supplemental Materials for other comparisons within this model).

Given (a) the observed positivity bias, where moral updating was stronger in Immoral-to-Moral scenarios than in Moral-to-Immoral scenarios, and (b) that interpreting reframing information as situational (vs. dispositional) increased positive moral updating, we tested whether situational interpretations mediate the effect of reframing direction on moral updating. A Bayesian multilevel model was used to estimate this mediation. The mean estimated total effect of reframing direction on moral updating was $b = 2.36$ [95% Bayesian credible interval = 1.96, 2.72], and the mean estimated direct effect was $b = 2.24$ [1.87, 2.63]. The mean estimated indirect effect of reframing direction on moral updating through informational judgments was $b = 0.11$ [0.07, 0.16], representing a 4.81% [2.80%, 6.81%] mediation. These results indicate that a small portion of the effect of reframing direction on moral updating is explained by informational judgments. It should be noted, however, that the causal effect of the mediator cannot be determined in the current design, as situational vs. dispositional interpretations were not themselves manipulated, and alternative models could explain the positivity bias in moral updating; further research is required to test

[3] A sensitivity analysis (using *simr*, Green & MacLeod, 2016) indicated that the 3-way interaction between Initial Condition, Reframing Condition, and Timepoint could be detected at a minimum effect size 75% below the observed effect size, while retaining ~80% power (Arend & Schäfer, 2019; Bloom, 1995). All fixed effects in the model were multiplied by 0.25, and a Monte Carlo simulation was used to conduct a z-test on the interaction term (*power* = 87.00% [78.80%, 92.89%], 100 simulations, function call: powerSim(model, nsim = 100, test = fixed("Initial:Reframing:Timepoint", method = "z"), seed = 123)).

[4] This slight positive shift in ratings in Immoral–Control scenarios was also observed in Study 1; one possibility is that, while the control information was intended to be morally irrelevant, some control segments may nonetheless have been perceived as slightly moral (e.g., visiting one's parents). This positive update may have been restricted to Immoral–Control (vs. Moral–Control) scenarios because Initial Immoral segments were highly valenced, and thus ensuing mundanely moral information may have had more of an impact on impressions.

[5] A sensitivity analysis (using *simr*, Green & MacLeod, 2016) indicated that the 3-way interaction between Initial Condition, Reframing Condition, and Timepoint could be detected at a minimum effect size 50% below the observed effect size, while retaining ~90% power. All fixed effects in the model were multiplied by 0.5, and a Monte Carlo simulation was used to conduct a z-test on the interaction term (*power* = 93.00% [86.11%, 97.14%], 100 simulations, function call: powerSim(model, nsim = 100, test = fixed("Initial:Reframing: Timepoint", method = "z"), seed = 123)).

[6] A sensitivity analysis (using *simr*, Green & MacLeod, 2016) indicated that the 2-way interaction between informational judgments and Reframing Condition could be detected at a minimum effect size 70% below the observed effect size, while retaining ~80% power. All fixed effects in the model were multiplied by 0.3, and a Monte Carlo simulation was used to conduct a z-test on the interaction term (*power* = 85.00% [76.47%, 91.35%], 100 simulations, function call: powerSim(model, nsim = 100, test = fixed("Informational: Reframing", method = "z"), seed = 123)).

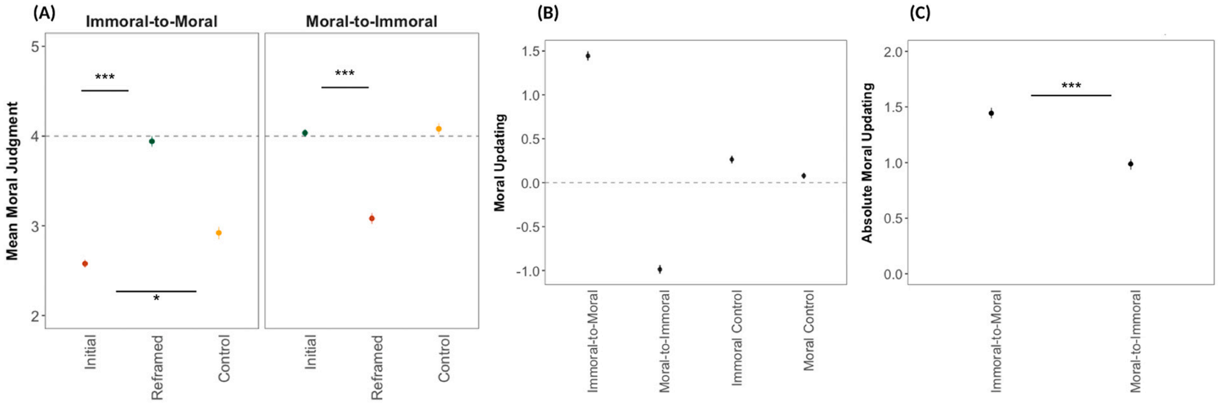

Fig. 3. Study 2 mean moral judgments and updating. A: Mean moral judgment for each condition, collapsed across act-based and person-based moral judgments. Error bars represent 95% confidence intervals. B: Difference between second-pass and first-pass moral judgments, for each scenario type. C: Magnitude of moral updating. There was greater updating for Immoral-to-Moral reframing relative to Moral-to-Immoral reframing. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ (after correction for multiple comparisons).

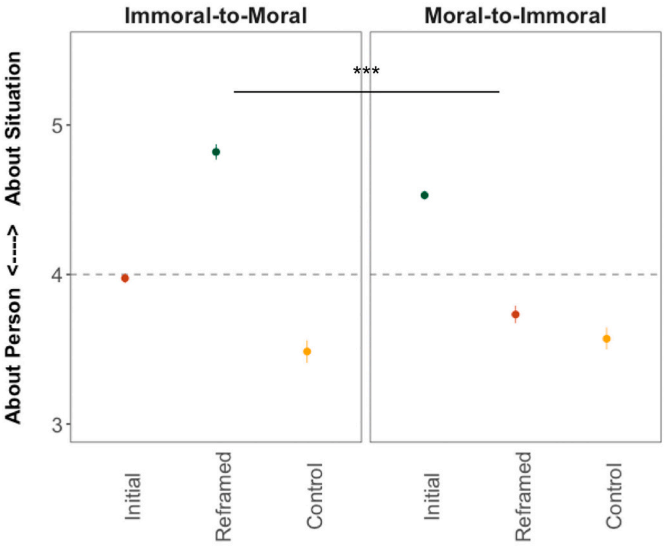

Fig. 4. Study 2 condition means for informational judgments. Moral information was rated as providing more information about the situation (relative to the scale midpoint), both when it was presented in the initial segment and in the reframing segment. Moral reframing information was rated as providing more information about the situation, compared to immoral reframing information. Error bars represent 95% confidence intervals. ***: $p < 0.001$.

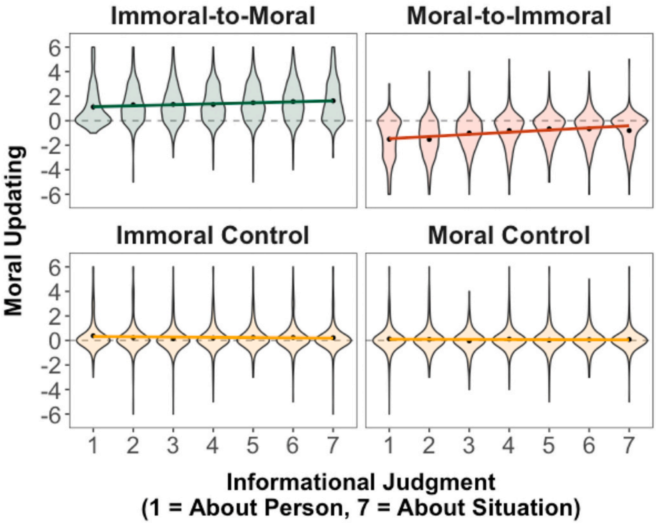

Fig. 5. Relationship between second-pass informational judgments and moral updating for each scenario type in Study 2. For Immoral-to-Moral scenarios and Moral-to-Immoral scenarios, rating reframing information as providing more information about the situation was associated with more positive (less negative) moral updating. There was no significant relationship between informational judgments and moral updating for Immoral–Control scenarios, or for Moral–Control scenarios.

whether situational vs. dispositional interpretations causally affect updating, and to test alternative models (Fiedler, Harris, & Schott, 2018; Pirlott & MacKinnon, 2016). In sum, the current results reveal a correlation, such that the positivity bias in moral updating may be partially explained by the interpretation of reframing information as situational.

3.3. Discussion

Study 2 was a preregistered, direct replication of both Study 1 and an initial exploratory follow-up study to it (reported in Supplemental Materials as Study 1.5). Replicating Study 1, an asymmetry was observed that was consistent with a positivity bias, such that moral updating was greater for Immoral-to-Moral scenarios, compared to Moral-to-Immoral scenarios. Further, additional effects were observed, consistent with initial exploratory results: first, moral reframing information was interpreted as more situational, compared to immoral reframing information; second, informational judgments and moral updating were correlated, such that in both Immoral-to-Moral and Moral-to-Immoral scenarios, situational interpretations of reframing information were associated with more positive moral updating; and finally, the observed positivity bias was partially mediated by situational interpretations. Observing these effects in a planned replication gives confidence that the results are robust.

Nonetheless, the specific mechanism underlying the observed positivity bias remains unclear. As discussed above, moral updating may depend on the relationship between new and old information. If new information is *directly related* to a previous behavior—as it is in the current *reframing* paradigm—then it may elicit more positive moral judgments by encouraging an interpretation of the agent's *past* behavior as being subject to situational constraints. By contrast, if the new information describes an *unrelated*, new behavior—as in an *addition* paradigm (Mann & Ferguson, 2015)—then the *new* behavior may be perceived as arising from a non-dispositional cause, which may lead to less extreme moral updating, but not a positivity bias. To test the relative impact of *reframing* and *addition* paradigms on moral updating, Study 3 compared attributions and moral updating following new information that either reframed past behavior or was irrelevant to it.

4. Study 3

Study 3 was a pre-registered (https://osf.io/jk9zy) experiment, comparing attributions and moral updating when participants received either reframing information or additional irrelevant information. The experiment used a 2 (*Paradigm Condition: Reframing* vs. *Addition*) x 2 (*Valence Direction: Moral-to-Immoral* vs. *Immoral-to-Moral*) mixed design, where Paradigm Condition was manipulated between-subjects, and Valence Direction was manipulated within-subjects.

As in Studies 1–2, in the Reframing condition, participants read one story about each target person, which evolved from a relatively moral (tragic) frame to a relatively immoral (taboo) frame, or vice versa. In the Addition condition, participants read a sequence of two unrelated stories about each target person—the relatively moral frame of one story, followed by the relatively immoral frame of a separate story, or vice versa. All of the scenarios that were presented in the Reframing condition were presented in the Addition condition (after recombining segments), meaning that stimulus features were controlled across the two paradigm conditions.

This experiment also addressed other limitations of Studies 1–2. For one, participants were previously asked to make *informational judgments* about whether a story segment provided relatively more information about the person or about the situation. It was assumed that these judgments reflected *causal attributions* of the agent's past behavior to the agent or to the situation, but the measure itself cannot support such a rich interpretation on its own. Further, recent work disputes the utility of the dispositional–situational distinction, and suggests that people's causal attributions are better characterized by two dimensions with independent explanatory power: *Externality*, or whether the cause is external (vs. internal) to the person, and *Stability*, or whether the cause is stable (vs. unstable) over time (Körner, Moritz, & Deutsch, 2020). In Study 3, these attributions were assessed by the *locus of causality* subscale and the *stability of cause* subscale from the Revised Causal Dimension Scale (CDSII; McAuley, Duncan, & Russell, 1992).

4.1. Hypotheses

Study 3 aimed to replicate and extend previous findings. In both the Reframing and Addition paradigm conditions, participants received two counter-valenced story segments about each target person. What differed between paradigm conditions was whether the story segments are about the same behavior (i.e., Reframing), or unrelated behaviors (i.e., Addition). The Reframing condition is similar to recent work on reinterpretation (Ferguson, Mann, Cone, & Shen, 2019; Mann & Ferguson, 2015), as the reframing segment can shed new light on a past behavior. The Addition condition is similar to past work on impression updating that has observed a negativity bias (Cone & Ferguson, 2015; Mende-Siedlecki, Baron, & Todorov, 2013; Reeder & Coovert, 1986) in that it presents a sequence of unrelated behaviors. Including both types of paradigms in the same experiment (as done in Mann & Ferguson, 2015) provided an opportunity to better characterize when and how the negativity bias can be overcome. Further, another objective of Study 3 was to better characterize the correlation observed in Study 2 between situational interpretations and moral updating, as the informational measure conflates multiple factors; to address this, informational judgments, externality judgments, and stability judgments were compared in the same experiment.

Our main hypotheses were as follows: (1) in the Reframing condition, consistent with the positivity bias observed in Studies 1–2, moral updating will be greater for Immoral-to-Moral targets (vs. Moral-to-Immoral), whereas in the Addition condition, consistent with a negativity bias, moral updating will be greater for Moral-to-Immoral targets (vs. Immoral-to-Moral); (2) new moral information will be rated as providing more situational information than new immoral information; (3) in the Reframing condition, there will be a positive correlation between situational interpretations and moral updating, while in the Addition condition, there will be a negative correlation; and (4) informational judgments will partially mediate moral updating. Further, we planned to: (1) explore whether moral and immoral information differ in externality judgments and stability judgments; (2) test for correlations between these attribution judgments and moral updating within each paradigm condition; and (3) test whether these attribution judgments mediate moral updating.

4.2. Methods and materials

4.2.1. Participants

The planned sample size was 500 (250 per paradigm condition), and participants were excluded if they quit before completing the survey, failed an attention check, or reported post-task that they did not attend to the task. Participants were recruited through Amazon Mechanical Turk in exchange for payment. The final sample consisted of 514 adults (265 identified as male, 239 identified as female, 3 identified as non-binary/other; $M_{\text{Age}} = 42.53$ years, $SD_{\text{Age}} = 12.77$ years), after excluding 17 participants for failing the attention check. After collecting data from all 514 participants, we conducted our analyses without collecting additional data. See footnote 6 for a sensitivity analysis tailored to the mixed effects design for this study. The study was approved by the Boston College IRB, and each participant provided informed consent before beginning the survey. While analyses were conducted with a subset of measures and participants, all measures, manipulations, and exclusions are reported.

4.2.2. Stimuli

Of the 24 root scenarios from Studies 1–2, 8 root scenarios were used in Study 3. Only the moral and immoral segments of each scenario were used; control segments were excluded. This subset of 8 scenarios was chosen such that, on average, initial moral and immoral segments of the scenarios had equivalent ratings for situational information (as rated by participants in Study 2; Initial Moral: $M = 4.30$, $SE = 0.10$; Initial Immoral: $M = 4.14$, $SE = 0.10$; $t(14) = -1.13$, $p = 0.28$). For additional analyses testing this equivalency in Study 3, see Supplemental Materials. These moral and immoral segments were presented in both the Reframing and Addition paradigm conditions—where for the Reframing condition, the segments appeared in the same combinations as in Studies 1–2, but for the Addition condition, the segments were mixed in new combinations (see Supplemental Materials for counterbalancing details and full scenario text).

4.2.3. Procedure

Participants were randomly assigned to the Reframing condition ($N = 257$) or the Addition condition ($N = 256$). Each participant learned about 4 Moral-to-Immoral target persons, and 4 Immoral-to-Moral target persons, all of which were presented in a randomized order. Unlike in Studies 1–2, no Moral-Control or Immoral-Control scenarios were used, as primary comparisons were between paradigm conditions.

For the Reframing condition, the procedure largely followed that of Study 2. Participants were introduced to a target and presented with their initial dilemma, which was framed as relatively moral or relatively immoral (e.g., Fig. 1a–1d); participants then made their first-pass judgments. Following this, participants were presented with new information (e.g., Fig. 1e), which reframed the initial information as relatively immoral (in the Imoral-to-Mmoral condition) or relatively moral (in the Moral-to-Immoral condition). After this reframing, participants made their second-pass judgments.

For the Addition condition, the procedure was similar, but instead of reframing the initial scenario with a new segment (Fig. 1e), participants were introduced to a second scenario (i.e., Fig. 1a-d, from a new scenario), up to and including a new moral or immoral decision. The name of the target person was consistent between the first and second scenario, such that participants had the experience of reading two disconnected stories about the same person. It is worth noting that this necessarily means participants in the Addition condition read more sentences per target than participants in the Reframing condition (14 vs. 9, as it takes 7 sentences to describe a dilemma, and 2 sentences to reframe it).

In both paradigm conditions, participants made first-pass and second-pass judgments where they responded to 8 items, including 3 *locus of causality* (i.e., *Externality)* items (e.g., "Is this cause something that reflects an aspect of this person or an aspect of the situation?" 1 = reflects an aspect of this person, 7 = reflects an aspect of the situation; McAuley et al., 1992), and 3 *stability of cause* (i.e., *Stability*) items (e.g., "Is this cause permanent or temporary?" 1 = permanent, 7 = temporary; McAuley et al., 1992). The discriminant validity of these two attribution measures was verified (see Supplemental Materials); for all downstream analyses, mean scores were used for each measure. Following the attribution measures, participants made an informational judgment ("Based on what you know so far, have you learned more about <agent>, or about the situation?" 1 = only about <agent>, 7 = only about the situation) and a moral judgment ("Is <agent> a moral person?" 1 = not at all, 7 = completely). As act-based vs. person-based judgment type did not produce theoretically important differences in Studies 1–2, only person-based moral judgments were tested.

4.2.4. Analysis

All data and analysis code are available on OSF (see Open Practices). Linear mixed effects models predicted four dependent variables in separate models: the absolute magnitude of moral updating, informational judgments, externality judgments, and stability judgments (*lme4* package; Bates et al., 2014). Fixed effects included: Paradigm Condition (Reframing, Addition), Valence Direction (Moral-to-Immoral, Immoral-to-Moral), and their interaction; by-subject random slopes were modeled for Valence Direction (but removed if convergence could not be achieved).

For correlation analyses, linear mixed effects models predicted moral updating (second-pass minus first-pass moral judgments) while modeling the fixed effects of: Paradigm Condition, Valence Direction, second-pass informational/externality/stability judgments, and their interactions. By-subject random slopes were modeled for Valence Direction and informational/externality/stability judgments (but removed if convergence could not be achieved).

P-values for fixed effects were obtained via Satterthwaite's degrees of freedom method (*lmerTest* package; Kuznetsova et al., 2017). We report partial Eta-squared values as effect sizes for interaction terms (*effectsize* package; Ben-Schachar et al., 2020). Contrasts within the model were tested simultaneously using the *multcomp* package (Hothorn et al., 2016). Effect sizes (Cohen's *d*) for contrasts were estimated by dividing the mean difference by the square root of the summed variance components (Brysbaert & Stevens, 2018). All reported *p*-values are corrected for multiple comparisons using the Tukey method.

For mediation analyses within each paradigm condition, Bayesian multilevel models (*brms* package; Bürkner, 2017) tested whether second-pass informational/externality/stability judgments mediate the relationship between Valence Direction and moral updating. Default, uninformative priors were used, and all Rhat values were < = 1.01, suggesting that the models had converged.

4.3. Results

4.3.1. Moral updating

Study 3 was designed to test whether the positivity bias observed in Studies 1–2 would be affected when moral judgments were updated under an Addition paradigm, as opposed to the Reframing paradigm used in Studies 1–2. Consistent with our hypothesis that the positivity bias would be preserved in the Reframing paradigm, and that the negativity bias would be restored in the Addition paradigm, a significant 2-way interaction[7] between Paradigm Condition and Valence Direction was observed, *Estimate* = 0.54, *SE* = 0.16, $t(512.62) = 3.31$, $p = 0.001$, $\eta^2_p = 0.021[0.003, 0.051]$. Planned contrast analyses within each Paradigm Condition revealed that, in the Reframing condition, the absolute magnitude of moral updating was greater for Immoral-to-Moral targets ($M = 1.09$, $SE = 0.07$) compared to Moral-to-Immoral targets ($M = 0.79$, $SE = 0.07$), *Estimate* = 0.30, *SE* = 0.07, $t(770) = 4.07$, $p = 0.0001$, $d = 0.22[0.11, 0.33]$. By contrast, in the Addition condition, the absolute magnitude of moral updating was marginally greater for Moral-to-Immoral targets ($M = 1.25$, $SE = 0.10$) compared to Immoral-to-Moral targets ($M = 1.01$, $SE = 0.10$), *Estimate* = 0.24, *SE* = 0.14, $t(1026) = 1.70$, $p = 0.090$, $d = 0.11[-0.02, 0.23]$. Thus, the positivity bias remained present when new information continued and reframed the initial story (replicating Studies 1–2), but a marginal negativity bias was present when two independent stories were presented (Fig. 6).

4.3.2. Second-pass informational judgments

In both the Reframing condition and in the Addition condition, planned contrast analyses revealed that final moral segments provided more situational information than final immoral segments (Reframing: *Estimate* = 0.481, *SE* = 0.10, $t(256) = 4.79$, $p < 0.0001$, $d = 0.28[0.17, 0.40]$; Addition: *Estimate* = 0.74, *SE* = 0.09, $t(771) = 8.24$, $p < 0.0001$, $d = 0.45[0.35, 0.56]$; see Supplemental Materials for other comparisons within the model). Thus, final moral segments were seen as providing more information about the situation, regardless of paradigm (Fig. 7a).

4.3.3. Correlations between moral updating and second-pass informational judgments

Moral updating was modeled using second-pass informational judgments, Paradigm Condition, Valence Direction, and their interactions. Four contrast analyses were licensed by significant two-way interactions (Paradigm Condition x informational judgments: *Estimate* = −0.10, *SE* = 0.05, $t(1858.56) = -2.07$, $p = 0.038$, $\eta^2_p = 0.002[0, 0.009]$; Valence Direction x informational judgments: *Estimate* = −0.13, *SE* = 0.05, $t(1960.32) = -2.61$, $p = 0.009$, $\eta^2_p = 0.0035[0.0002, 0.0105]$). These analyses revealed significant positive correlations between informational judgments and moral updating among: Reframing targets (*Estimate* = 0.21, *SE* = 0.04, $z = 5.69$, $p < 0.001$, $d = 0.11[0.07, 0.15]$), Addition targets (*Estimate* = 0.31, *SE* = 0.03, $z = 9.22$, $p < 0.001$, $d = 0.16[0.13,0.20]$), Moral-to-Immoral targets (*Estimate* = 0.33, *SE* = 0.03, $z = 9.79$, $p < 0.001$, $d = 0.17[0.14, 0.20]$), and Immoral-to-Moral targets (*Estimate* = 0.20, *SE* = 0.04, $z = 5.17$, $p < 0.001$, $d = 0.10[0.06, 0.14]$) (Fig. 8a). This ran counter to our expectation that the correlation might reverse in the Addition condition, but suggests that the relationship holds more broadly.

4.3.4. Second-pass externality and stability judgments

For externality judgments, planned contrast analyses revealed that final moral segments elicited more *external* attributions than final immoral segments (Reframing: *Estimate* = 0.70, *SE* = 0.10, $t(256) = 6.65$, $p < 0.0001$, $d = 0.40[0.28, 0.51]$; Addition: *Estimate* = 1.03, *SE* = 0.09, $t(771) = 10.98$, $p < 0.0001$, $d = 0.62[0.51, 0.73]$; see Supplemental Materials for other comparisons). Thus, final moral information was more likely to elicit more external attributions, regardless of

[7] A sensitivity analysis (using *simr*, Green & MacLeod, 2016) indicated that the 2-way interaction between Paradigm Condition and Valence Direction could be detected at a minimum effect size 15% below the observed effect size, while retaining ~80% power (Arend & Schäfer, 2019; Bloom, 1995). All fixed effects in the model were multiplied by 0.85, and a Monte Carlo simulation was used to conduct a z-test on the interaction term (*power* = 80.30% [77.70%, 82.72%], 1000 simulations, function call: powerSim(model, nsim = 1000, test = fixed ("Paradigm:Valence", method = "z"), seed = 123)).

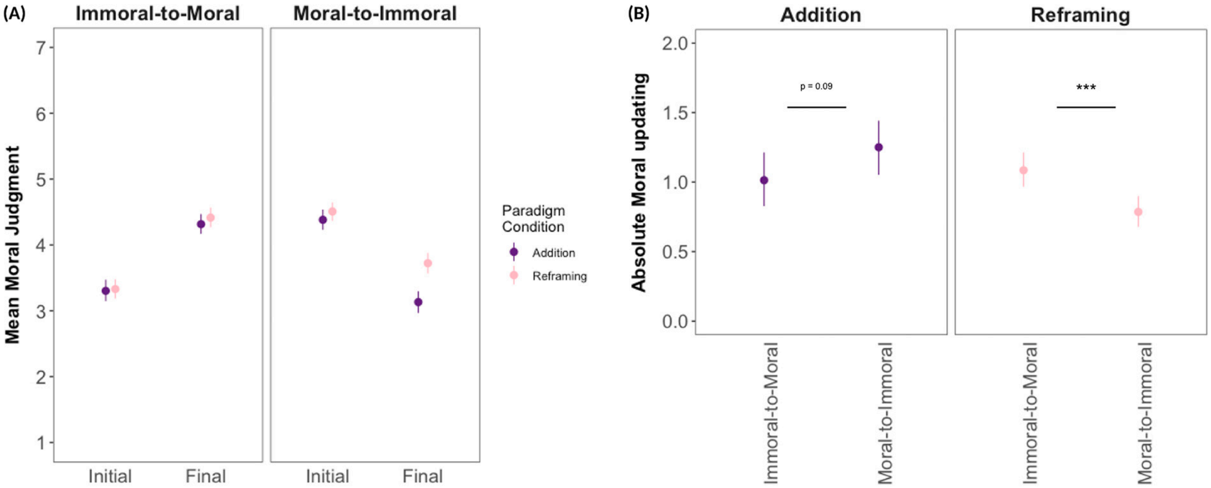

Fig. 6. Study 3 mean moral judgments and updating. A: Mean moral judgment for each condition. Error bars represent 95% confidence intervals. B: Magnitude of moral updating. There was a marginal negativity bias in the Addition condition and a significant positivity bias in the Reframing condition. ***: $p < 0.001$.

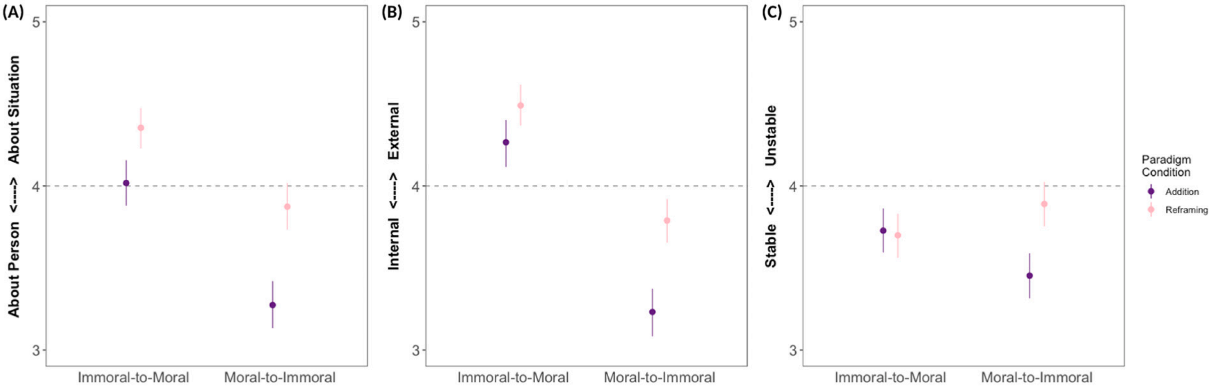

Fig. 7. Study 3 condition means second-pass judgments by condition. A: informational judgments; B: externality judgments; C: stability judgments. Error bars represent 95% confidence intervals.

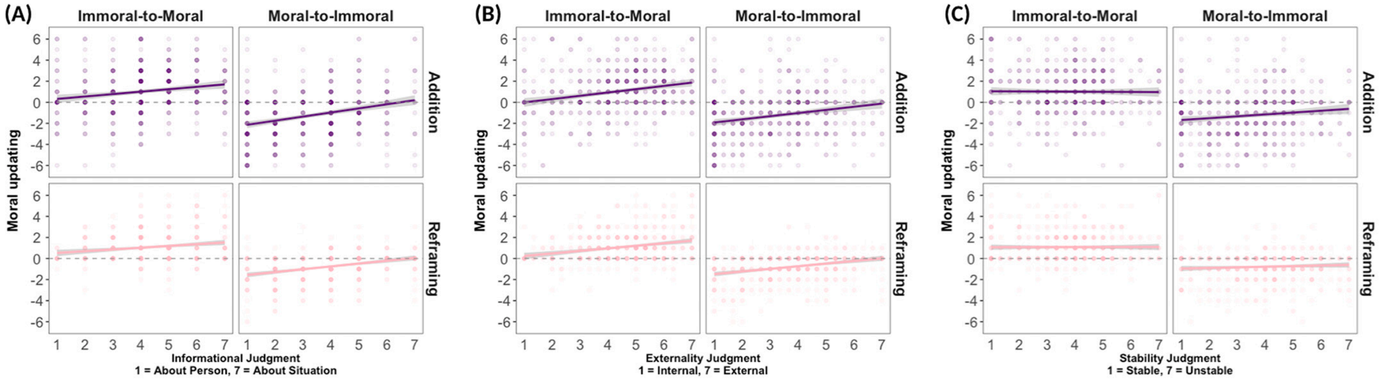

Fig. 8. Correlations between second-pass judgments and moral updating in Study 3. A: Relationship between second-pass informational judgments and moral updating for each condition. B: Relationship between second-pass externality judgments and moral updating for each condition. C: Relationship between second-pass stability judgments and moral updating for each condition.

paradigm (Fig. 7b).

For stability judgments, there was a 2-way interaction between Paradigm Condition and Valence Direction (*Estimate* = −0.46, *SE* = 0.12, $t(1541.43) = -3.89, p < 0.001, \eta^2_p = 0.010[0.002, 0.021\ 1]$). Here, a new pattern was observed: in the Reframing condition, final *immoral* segments elicited more unstable attributions than final moral segments (*Estimate* = 0.19, *SE* = 0.09, $t(256) = 2.15$, $p = 0.032$, $d = 0.12[0.01, 0.23]$); in the Addition condition, final *moral* segments elicited more unstable attributions than final immoral segments (*Estimate* = 0.27, *SE* = 0.08, $t(771) = 3.35$, $p < 0.001$, $d = 0.17[0.07, 0.28]$; see Supplemental Materials for other comparisons) (Fig. 7c).

4.3.5. Correlations between moral updating and second-pass externality and stability judgments

A potential confound in the correlations between moral updating and informational judgments reported above (and in Study 2) is that the informational measure (person vs. situation) conflates multiple factors related to dispositional–situational attributions: *Externality* (whether the cause is internal or external to the person), and *Stability* (whether the cause is stable or unstable over time; Körner et al., 2020). To address this, we estimated correlations between moral updating and externality judgments, and between moral updating and stability judgments.

For externality judgments, no three-way or two-way interactions were significant (see Supplemental Materials for statistics), but for comparison with the above analyses of informational judgments, we examined the equivalent four contrasts. Here, the pattern of correlations replicated the pattern seen for informational judgments (Fig. 8b). There were significant positive correlations between externality judgments and moral updating among: Reframing targets ($Estimate = 0.24$, $SE = 0.04$, $z = 6.53$, $p < 0.001$, $d = 0.14[0.11, 0.18]$), Addition targets ($Estimate = 0.31$, $SE = 0.03$, $z = 9.16$, $p < 0.001$, $d = 0.16[0.13, 0.19]$), Moral-to-Immoral targets ($Estimate = 0.27$, $SE = 0.03$, $z = 8.03$, $p < 0.001$, $d = 0.16[0.13, 0.19]$), and Immoral-to-Moral targets ($Estimate = 0.27$, $SE = 0.04$, $z = 7.50$, $p < 0.001$, $d = 0.12[0.09, 0.16]$). The results here suggest that attributions to external causes may contribute to the observed correlation between our informational measure and moral updating.

For stability judgments (scale anchors: 1 = stable, 7 = unstable), there was a significant 2-way interaction with Valence Direction ($Estimate = -0.11$, $SE = 0.05$, $t(1870.48) = -2.08$, $p = 0.038$, $\eta^2_p = 0.002[0, 0.009]$; see Supplemental Materials for other statistics). As above, we performed four contrast analyses. Here, the pattern of correlations differed from the pattern seen for informational judgments and for externality judgments (Fig. 8c). For Reframing targets, there was no correlation between stability judgments and moral updating ($Estimate = 0.02$, $SE = 0.04$, $z = 0.62$, $p = 0.923$, $d = 0.01[-0.02,0.05]$); for Addition targets, there was a marginally positive correlation between stability judgments and moral updating, such that more *unstable* ratings predicted more positive moral updating ($Estimate = 0.09$, $SE = 0.04$, $z = 2.48$, $p = 0.064$, $d = 0.05[0.01, 0.08]$); for Moral-to-Immoral targets, there was a significant positive correlation ($Estimate = 0.11$, $SE = 0.04$, $z = 3.08$, $p = 0.012$, $d = 0.06[0.02, 0.09]$); and for Immoral-to-Moral targets, there was no correlation ($Estimate = 0.002$, $SE = 0.038$, $z = 0.050$, $p = 0.999$, $d = 0.001[-0.036, 0.038]$). Thus, when examining two potential component factors of our informational measure, it appears that judgments of *Externality* are more consistent with the pattern observed for informational judgments, rather than judgments of *Stability*. This suggests that the pattern of correlations observed for informational judgments has more to do with excusing prior behavior on the basis of revealed external causes, as opposed to excusing it on the basis of revealed unstable causes.

4.3.6. Mediation analyses

To confirm that moral updating was driven by attributions to external causes and by situational interpretations (as opposed to attributions to unstable causes), we performed a mediation analysis, as in Study 2. Within the Reframing condition and within the Addition condition, we tested whether informational judgments, externality judgments, and stability judgments mediate the relationship between Valence Direction and moral updating.

In the Reframing condition, informational and externality judgments partially mediated the effect of Valence Direction on moral updating, while a negligible mediation was observed for stability judgments. For informational judgments, the mean estimated indirect effect was $b = 0.11$ [0.06, 0.16], representing a 5.71%[3.28%, 8.14%] mediation (total effect: $b = 1.87$ [1.70, 2.03]; direct effect: $b = 1.77$[1.60, 1.91]). For externality judgments, the mean estimated indirect effect was $b = 0.17$ [0.12, 0.23], representing a 9.24% [6.22%, 12.27%] mediation (total effect: $b = 1.87$ [1.71, 2.04], direct effect: $b = 1.70$ [1.54, 1.87]). By contrast, for stability judgments, the mean estimated indirect effect was $b = -0.006$ [−0.020, 0.005], representing a − 0.31% [−0.99%, 0.36%] mediation (total effect: $b = 1.87$ [1.70, 2.04], direct effect: $b = 1.88$ [1.71, 2.04]).

In the Addition condition, the same general pattern was observed: informational and externality judgments partially mediated the effect of Valence Direction on moral updating, but a negligible mediation was observed for stability judgments. For informational judgments, the mean estimated indirect effect was $b = 0.24$ [0.15, 0.32], representing a 10.40%[6.68%, 14.12%] mediation (total effect: $b = 2.26$ [2.00, 2.53], direct effect: $b = 2.03$[1.75, 2.29]). For externality judgments, the mean estimated indirect effect was $b = 0.32$ [0.22, 0.42], representing a 14.00%[9.59%, 18.41%] mediation (total effect: $b = 2.26$ [1.99, 2.52], direct effect: $b = 1.95$ [1.67, 2.21]). Again, by contrast, for stability judgments, the mean estimated indirect effect was $b = 0.025$ [−0.001, 0.056], representing a 1.11%[−0.15%, 2.38%] mediation (total effect: $b = 2.27$ [1.98, 2.53], direct effect: $b = 2.24$ [1.98, 2.53]). Thus, in both the Reframing and Addition conditions, informational and externality judgments appear to partially mediate moral updating, while stability ratings do not. The correlation patterns (for informational and externality judgments) were unexpected, considering that a positivity bias in moral updating was present in the Reframing condition, but not in the Addition condition. This suggests that the same underlying mechanism may be at play in both paradigm conditions, where more positive moral updating is predicted by more external attributions following new information, and by more situational interpretations of new information.

4.4. Discussion

One objective of Study 3 was to replicate key results in Studies 1–2 and to extend them by contrasting updating that occurs in a reframing paradigm (i.e., contextualizing old information with new) with updating that occurs in an addition paradigm (i.e., adding new, irrelevant, counter-valenced information). Including both paradigms in the same experiment helped characterize when and how the negativity bias can be overcome. In Study 3, we replicated the previously observed positivity bias in moral updating in the Reframing condition, but also observed a marginal negativity bias in the Addition condition, consistent with prior work using addition paradigms (Cone & Ferguson, 2015; Mende-Siedlecki, Baron, & Todorov, 2013; Reeder & Coovert, 1986). These findings underscore the importance of reframing for moral updating in the positive direction.

A second objective of Study 3 was to better characterize the correlation between moral updating and informational judgments (i.e., judgments of whether new information says more about the person, or about the situation). The informational measure conflates multiple factors related to attributions, but these factors can be brought into focus by collecting separate judgments of the locus of the cause (*Externality*), and the stability of the cause (*Stability*). We found that new moral information elicited more situational interpretations and more external attributions, and both of these judgments predicted more positive updating. In contrast, this pattern of correlations did not replicate for stability judgments. Likewise, both informational judgments and externality judgments mediated moral updating, while stability judgments did not. Thus, when it comes to causal attributions, it seems that the perceived externality of the cause plays a more consistent role in updating, than the perceived stability of the cause. The independence of these two types of causal attribution is in line with recent work demonstrating the importance of de-confounding the dispositional–situational dichotomy in attributions (Körner et al., 2020).

Furthermore, somewhat surprisingly, these patterns of correlation and mediation (for informational and externality judgments) were present in both the Reframing and Addition conditions, suggesting that the relationship might hold across different types of paradigms.

5. General discussion

The present work examined moral updating after decisions in moral dilemmas were reframed, either by revealing a hidden economic incentive behind a seemingly selfless decision, or revealing a hidden positive outcome of a seemingly selfish decision. Findings from three studies extend our understanding of explicit moral updating in several ways. First, Study 1 showed that, in a reframing paradigm, the negativity bias (Baumeister et al., 2001; Rozin & Royzman, 2001) can be reversed: moral judgments became more positive in Immoral-to-Moral scenarios than they became negative in Moral-to-Immoral scenarios, despite participants having access to the same information after reframing. Study 2 replicated this unanticipated "positivity bias" and also revealed an asymmetry in the qualitative interpretation of the reframing information: moral reframing information was rated as providing more situational information, and, across scenario types, more situational interpretations were associated with more positive moral updating. Further, the effect of reframing direction on the moral updating was partially mediated by situational interpretations. Finally, Study 3 replicated the positivity bias observed after reframing (as in Studies 1–2), but also produced a marginal negativity bias in an addition paradigm, where independent, counter-valenced information was appended to a story. This suggests that the relevance of novel information aimed to update moral judgment matters, and might change how exculpatory information is interpreted. Study 3 also clarified the mediation by situational interpretations, showing that perceived locus of causality (i.e., as external vs. internal) plays a more consistent role in updating than perceived stability of cause (i.e., stable vs. unstable).

5.1. *Both increasing and decreasing certainty may contribute to moral updating*

In the present work, moral information was operationalized as information about a tragic dilemma. By design, tragic dilemmas involve a difficult decision (Tetlock, et al., 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 more external attributions and positive updating: the initial taboo decision *appeared* selfish and immoral because of participants' limited understanding of the complicated context; recognizing the difficulty of the decision may help move moral judgments.

This mechanism is subtly different from the one that may be occurring in the implicit updating work of Mann and Ferguson (2015). In that paradigm, a man breaks into two houses and takes "precious things"; but later, it is revealed that he broke into the houses because they were on fire, and that the precious things were young children. As the agent's motive is underspecified at first, participants' initial evaluations may rely heavily on the agent's actions alone. Thus, when the new information (the nature of the precious things) is revealed, it can wholly explain the prior behavior, and may prompt a complete undoing of initial evaluations. That is, participants may move toward a state of greater certainty, as the agent's initial motive—underspecified but assumed to be negative—gets replaced by an indisputably positive one. In contrast, the present work presents motives that are potentially coexisting—the revelation of the second motive cannot erase the first one. In Immoral-to-Moral scenarios, participants first learn about a selfish, immoral motive in the context of a taboo dilemma, then learn about the existence of a relatively selfless, moral motive in the context of a tragic dilemma. In this case, the inherent ambiguity and difficulty of tragic dilemmas may move participants toward a state of greater *uncertainty*, which destabilizes their initial evaluation, and prompts a consideration of the broader context.

We tentatively suggest that there are at least two routes to reframing an initially negative impression: (1) completely reinterpreting the old information because the new information demands it (e.g., by providing a new motive that replaces the old one), and (2) weakening the tie between the old information and the impression because new information suggests there are more variables to consider (e.g., co-existing motives). Both processes are distinguishable from learning about two unrelated pieces of information (e.g., one behavior with a good motive, another behavior with a bad motive). Future work might compare the two routes, both within an implicit paradigm and within an explicit paradigm. For example, it may be that method (2), which introduces uncertainty and associated metacognitive processing (e.g., reflecting on one's own uncertainty in this case), will change explicit judgments more than implicit ones.

5.2. *Initially negative impressions can be updated with relevant new evidence*

An initial negative impression can be updated via reframing, but it can also be updated by observing a particular behavior evolve over time. In one set of studies, participants observed someone choose between taking money for themselves or sending painful shocks to someone else (Siegel, Mathys, Rutledge, & Crockett, 2018). In this paradigm, beliefs about bad agents (who harmed others for personal gain) were more uncertain and more rapidly updated than beliefs about good agents. That is, beliefs about bad agents were more volatile than beliefs about good agents. These findings are in line with the idea that negative judgments can updated if *relevant* new evidence is provided—in this case, new instances of a specific behavior. By contrast, in addition paradigms (e.g., as in Study 3), new evidence comes from completely unrelated behaviors and may not update initial negative judgments. Why exactly unrelated counterevidence fails to induce moral updating could stem from a number of reasons: it may be that the original immoral action was more threatening, or it may be that people do not form generalizable impressions of someone as good or bad "on the whole", and that assessments are more context-dependent. Future work is necessary to disentangle these hypotheses.

5.3. *Implications*

Our findings may have potential implications for how criminal defendants are perceived by juries when new information comes to light. Information about a criminal defendant's moral behavior might be considered immaterial to a criminal charge (Davies, 1991), but jurors are also expected to be sensitive to new evidence that provides more context for the criminal act itself. Our work suggests that, if the new information is relevant to prior behavior, it may correct a bad first impression. At the same time, our work 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).

5.4. *Future directions*

Much of the prior work on updating negative impressions has examined implicit, rather than explicit, evaluations of targets (Mann, Cone, Heggeseth, & Ferguson, 2019; Mann & Ferguson, 2015, 2017; Wyer, 2010). These findings suggest that implicit evaluations are more easily updated than dual-systems accounts of cognition would suggest, especially when new evidence is diagnostic, believable, and/or allows for reinterpretation of past information (Ferguson et al., 2019; Mann et al., 2019; Mann & Ferguson, 2015, 2017; Wyer, 2010). The current work departs from these studies by probing updates to explicit impressions of targets instead. While we hope the employment of these measures will extend our overall understanding of impression updating, the present work does not directly compare explicit vs. implicit impressions; it may be particularly informative to compare the two measures when the new evidence is weaker (Ferguson et al., 2019). The moral dilemmas presented in the current study, which are inherently ambiguous, may cause explicit and implicit updating to diverge for a couple reasons. For one, moral reframing information may have moved participants toward a state of greater uncertainty about the agent's motives, perhaps prompting metacognitive processing of one's own uncertainty; this form of uncertainty may affect explicit updating more than implicit updating. Second, in any explicit updating paradigm, there is an implied social (experimenter-as-audience) context (Ferguson et al., 2019); participants may have assumed that the experimenters intended for moral reframing information to lead to greater reinterpretation, whereas in an implicit updating paradigm, the positivity bias observed here may be weakened due to lack of access to such assumptions. These two possibilities, and the effect of different types of social context (experimenter-as-audience and otherwise) on updating in general, merit detailed exploration in future work.

The present work also focused on moral judgments, as opposed to overall evaluations of people (Anderson, 1965; Baumeister et al., 2001; Mann & Ferguson, 2015, 2017; Rozin & Royzman, 2001). Although mounting evidence suggests that moral character dominates overall impressions (Brambilla et al., 2019; Goodwin, 2015), moral impressions may not always track overall impressions. For instance, in moral dilemmas, reframing information may update moral impressions without updating an overall impression of the person. This potential decoupling is worth investigating further, by directly comparing global vs. moral impressions.

Individuals have also been found to differ in the degree to which they make correspondent inferences (overly dispositional attributions for others' behaviors, even when behaviors are highly constrained by situational factors; Scopelliti, Min, McCormick, Kassam, & Morewedge, 2018). This individual difference can be assessed using the Neglect of External Demands scale, and is distinct from more general measures including cognitive ability and cognitive reflection (Scopelliti et al., 2018). Future work may explore the potential moderating role of this individual difference measure in moral updating following reframing. People who tend to be overconfident in their dispositional attributions may be less likely to exhibit the positivity bias, perhaps because they will discount reframing information overall, regardless of valence.

Finally, the role of uncertainty in moral updating merits further investigation. As discussed above, more external attributions following reframing information may indicate a move toward a state of greater uncertainty, as the observer considers the force of the extra motive afforded by the situation (e.g., protecting workers' livelihoods). Future work might: (1) collect confidence judgments alongside moral judgments to examine the importance of uncertainty for moral updating; and (2) influence risk preferences by forcing people to act on their beliefs (e.g., via partner choice decisions or decisions to trust). People may be more flexible in their belief updating depending on the type of social decision they are asked to make.

6. Conclusion

The present work compared explicit moral updating following contextualized moral-to-immoral and immoral-to-moral reframing of moral dilemmas. We found that the negativity bias in updating can be reversed when new information reframes earlier information, and that the negativity bias is partially maintained when additional, independent evidence is added. The "positivity bias" following reframing was partially explained by the extent to which reframing information elicits external causal attributions. Given the present results, and results from prior work, we speculate that reframing information may promote updating by increasing *or* decreasing the perceiver's certainty in their initial judgment. Future research on moral updating may benefit from a sensitivity to such qualitative features of new information.

Open practices

All data and analysis code for the paper are available on the Open Science Framework (https://osf.io/3cyaj/?view_only=9c5849eee24d4b24b1c106802893f59e).

Acknowledgments

This work was supported by the John Templeton Foundation, the National Science Foundation, the Roux Institute and the Harold Alfond Foundation, and the Boston College Research Incentive Grant. The authors would like to thank: the Morality Lab, the Implicit Social Cognition Lab, the Reasoning in Moral Thought and Action Symposium, Kevin Jiang, Justin Martin, BoKyung Park, the editor, and the reviewers for their feedback.

Conflicts of interest

None declared.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2022.104310.

References

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology, 70*(4), 394.

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods, 24*(1), 1.

Atran, S., & Axelrod, R. (2008). Reframing sacred values. *Negotiation Journal, 24*(3), 221–246.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition, 108*(2), 381–417.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv preprint. arXiv:1406.5823.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370.

Ben-Schachar, M. S., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software, 5*(56), 2815.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review, 19*(5), 547–556.

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology, 82*, 64–73.

Brannon, S. M., & Gawronski, B. (2018). In search of a negativity bias in expectancy violation. *Social Cognition, 36*(2), 199–220.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1).

Bürkner, P. C. (2017). *Advanced Bayesian multilevel modeling with the R package brms*. arXiv:1705.11123.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015.

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*(1), 37.

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Can we change our implicit minds? New evidence for how, when, and why implicit impressions can be rapidly revised. *Advances in Social Psychology, 56*, 131–199.

Critcher, C., Helzer, E., Tannenbaum, D., & Pizarro, D. (2012). Actions speak less loud than sentiments: A new model of moral judgment. In Z. Gürhan-Canli, C. Otnes, & R. Zhu (Eds.)*, Vol. 40. NA- advances in consumer research* (pp. 125–128). Association for Consumer Research.

Davies, S. M. (1991). Evidence of character to prove conduct: A reassessment of relevancy. *Criminal Law Bulletin, 27*, 504–524.

Driver, J. (2006). Virtue theory. In J. Dreier (Ed.), *Contemporary debates in moral theory* (pp. 113–120). Blackwell.

Faul, F., Erdfelder, E., Lang, A., & 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.

Fein, S. (1996). Effects of suspicion on attributional thinking and the correspondence bias. *Journal of Personality and Social Psychology, 70*(6), 1164.

Fein, S., Hilton, J. L., & Miller, D. T. (1990). Suspicion of ulterior motivation and the correspondence bias. *Journal of Personality and Social Psychology, 58*(5), 753.

Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and how implicit first impressions can be updated. *Current Directions in Psychological Science, 28*(4), 331–336.

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests–an analysis of articles published in 2015. *Journal of Experimental Social Psychology, 75*, 95–102.

Fiske, A. P., & Tetlock, P. E. (1997). Taboo trade-offs: Reactions to transactions that transgress the spheres of justice. *Political Psychology, 18*(2), 255–297.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology, 38*(6), 889–906.

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology, 15*(1), 183–217.

Gawronski, B., & Brannon, S. M. (2019). What is cognitive consistency and why does it matter? In E. Harmon-Jones (Ed.), *Cognitive dissonance: Progress on a pivotal theory in social psychology* (2nd ed.). Washington, DC: American Psychological Association.

Gilbert, D. T., & Jones, E. E. (1986). Perceiver-induced constraint: Interpretations of self-generated reality. *Journal of Personality and Social Psychology, 50*(2), 269.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*(1), 21.

Ginges, J., Atran, S., Medin, D., & Shikaki, K. (2007). Sacred bounds on rational resolution of violent political conflict. *Proceedings of the National Academy of Sciences, 104*(18), 7357–7360.

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science, 24*(1), 38–44.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464.

Hanselmann, M., & Tanner, C. (2008). Taboos and conflicts in decision making: Sacred values, decision difficulty, and emotions. *Judgment and Decision making, 3*(1), 51–63.

Heider, F (1958). *The psychology of interpersonal relations*. John Wiley & Sons Inc.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., & Hothorn, M. T. (2016). *Package 'multcomp'. Simultaneous inference in general parametric models*. Vienna, Austria: Project for Statistical Computing.

Kim, M. J., Mende-Siedlecki, P., Anzellotti, S., & Young, L. (2021). Theory of mind following the violation of strong and weak prior beliefs. *Cerebral Cortex, 31*(2), 884–898.

Klein, N., & O'Brien, E. (2016). The tipping point of moral change: When do good and bad acts make good and bad actors? *Social Cognition, 34*(2), 149–166.

Körner, A., Moritz, S., & Deutsch, R. (2020). Dissecting dispositionality: Distance increases stability of attribution. *Social Psychological and Personality Science, 11*(4), 446–453.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(1), 1–26.

Levine, S., Mikhail, J., & Leslie, A. M. (2018). Presumed innocent? How tacit assumptions of intentional structure shape moral judgment. *Journal of Experimental Psychology: General, 147*(11), 1728.

Lichtenstein, S., Gregory, R., & Irwin, J. (2007). What's bad is easy: Taboo values, affect, and cognition. *Judgment and Decision Making, 2*(3), 169–188.

Mandel, D. R., & Vartanian, O. (2008). Taboo or tragic: Effect of tradeoff type on moral choice, conflict, and confidence. *Mind & Society, 7*(2), 215–226.

Mann, T. C., Cone, J., Heggeseth, B., & Ferguson, M. J. (2019). Updating implicit impressions: New evidence on intentionality and the affect misattribution procedure. *Journal of Personality and Social Psychology, 116*(3), 349.

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108*(6), 823.

Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology, 68*, 122–127.

McAuley, E., Duncan, T. E., & Russell, D. W. (1992). Measuring causal attributions: The revised causal dimension scale (CDSII). *Personality and Social Psychology Bulletin, 18*, 566–573. https://doi.org/10.1177/0146167292185006

McGraw, A. P., & Tetlock, P. E. (2005). Taboo trade-offs, relational framing, and the acceptability of exchanges. *Journal of Consumer Psychology, 15*(1), 2–15.

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.

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology, 116*(2), 215.

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology, 66*, 29–38.

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

Reeder, G. D. (1993). Trait-behavior relations and dispositional inference. *Personality and Social Psychology Bulletin, 19*(5), 586–593.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86*(1).

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition, 4*(1), 1–17.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320.

Scopelliti, I., Min, H. L., McCormick, E., Kassam, K. S., & Morewedge, C. K. (2018). Individual differences in correspondence bias: Measurement, consequences, and correction of biased interpersonal attributions. *Management Science, 64*(4), 1879–1910.

Shirkey, H. B. (2010). Last attorney to the jury box is a rotten egg: Overcoming psychological hurdles in the order of presentation at trial. *Ohio State Journal of Criminal Law, 8*, 581.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour, 2*(10), 750–756.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52*(4), 689.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131.

Tadros, V., & Tierney, S. (2004). The presumption of innocence and the human rights act. *The Modern Law Review, 67*(3), 402–434.

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology, 47*(6), 1249–1254.

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

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences, 7*(7), 320–324.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & 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.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition, 126*(2), 326–334.

Van Zyl, L. (2007). Can virtuous people emerge from tragic dilemmas having acted well? *Journal of Applied Philosophy, 24*(1), 50–61.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistcal power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020.

Wyer, N. A. (2010). You never get a second chance to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Social Cognition, 28*(1), 1–19.