Cognition 239 (2023) 105570

Contents lists available at ScienceDirect

Cognition

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

Original articles

One strike and you're a lout: Cherished values increase the stringency of moral character attributions

Joshua Rottman [a,*], Emily Foster-Hanson [b], Sam Bellersen [c]

[a] Department of Psychology, Franklin and Marshall College, United States of America
[b] Department of Psychology, Swarthmore College, United States of America
[c] Department of Philosophy, Franklin and Marshall College, United States of America

ARTICLE INFO

Dataset link: https://osf.io/4yfuw

Keywords:
Moral character
Moral pluralism
True self
Negativity bias
Moral updating
Moral inconsistency

ABSTRACT

Moral dilemmas are inescapable in daily life, and people must often choose between two desirable character traits, like being a diligent employee or being a devoted parent. These moral dilemmas arise because people hold competing moral values that sometimes conflict. Furthermore, people differ in which values they prioritize, so we do not always approve of how others resolve moral dilemmas. How are we to think of people who sacrifice one of our most cherished moral values for a value that we consider less important? The "Good True Self Hypothesis" predicts that we will reliably project our most strongly held moral values onto others, even after these people lapse. In other words, people who highly value generosity should consistently expect others to be generous, even after they act frugally in a particular instance. However, reasoning from an error-management perspective instead suggests the "Moral Stringency Hypothesis," which predicts that we should be especially prone to discredit the moral character of people who deviate from our most deeply cherished moral ideals, given the potential costs of affiliating with people who do not reliably adhere to our core moral values. In other words, people who most highly value generosity should be quickest to stop considering others to be generous if they act frugally in a particular instance. Across two studies conducted on Prolific (N = 966), we found consistent evidence that people weight moral lapses more heavily when rating others' membership in highly cherished moral categories, supporting the Moral Stringency Hypothesis. In Study 2, we examined a possible mechanism underlying this phenomenon. Although perceptions of hypocrisy played a role in moral updating, personal moral values and subsequent judgments of a person's potential as a good cooperative partner provided the clearest explanation for changes in moral character attributions. Overall, the robust tendency toward moral stringency carries significant practical and theoretical implications.

1. Introduction

A typically honest person may occasionally tell a lie to help a friend avoid a painful truth. A vegetarian may sometimes eat meat as a gesture of respect when dining in others' homes. Every now and then, a philanthropist may refrain from donating her money and instead use it to indulge her children. People fall short in upholding their moral commitments, not only due to failures of self-control and sanguine reframings of personal wrongdoings (Batson, 2016; Shalvi et al., 2015; Valdesolo & DeSteno, 2007), but also because people prioritize different moral values in different situations. The moral domain is variegated (e.g., Flanagan, 2017; Graham et al., 2013; Sinnott-Armstrong & Wheatley, 2014), so there are many ways to be a moral person. Being a fair person, a brave person, or a caring person are each associated with divergent characteristics (Walker & Hennig, 2004), and when people are faced with conflicting moral demands, they must often sacrifice one moral commitment in order to uphold another discordant commitment (see Graham et al., 2015). For example, if a teacher who is typically equitable in their grading encounters a student enduring personal struggles, they may choose to grade more leniently, thus prioritizing compassion at the expense of fairness. In such a case, should this teacher still be considered to be a fair and equitable person? The answer to this question may vary based on people's subjective value hierarchies. That is, an observer's judgment about the teacher's moral character might depend on that observer's relative valuation of fairness as opposed to care. Here, we investigate how people's moral priorities shape their evaluations of others who sacrifice a typically upheld moral value in favor of a competing moral value.

The present research builds upon the observation that we care tremendously, for good reason, about others' moral tendencies. We

[*] Corresponding author.
E-mail address: jrottman@fandm.edu (J. Rottman).

https://doi.org/10.1016/j.cognition.2023.105570
Received 21 September 2022; Received in revised form 17 July 2023; Accepted 18 July 2023
Available online 1 August 2023

persistently categorize others based on their moral character (van Leeuwen et al., 2012), in part as a product of a fundamental drive to conceptualize the world as subdivided into discrete categories (Murphy, 2002). Beyond this general tendency, moral characteristics shape our perceptions of others more fundamentally than most other traits, including those that are diagnostic of competence or warmth (Brambilla & Leach, 2014; Goodwin et al., 2014), perhaps because moral traits are particularly diagnostic of whether a person is likely to be a helpful and trustworthy (or threatening) social partner (see Brambilla et al., 2021). Throughout our social interactions, we expend copious energy seeking and spreading information about others' moral tendencies (Boehm, 2012). Pursuing reliable knowledge about people's moral category membership (e.g., whether somebody is a caring person, a loyal person, or a selfless person) is adaptive: Determining who possesses values that we care about can help inform decisions about who to work with, befriend, or marry. But given that people rarely adhere to moral ideals in all circumstances, we must carefully decide when to discount moral failures. If we are too forgiving, we risk cooperating with people who will take advantage of us. If we are too intolerant, we risk having overly high expectations of others and losing out on worthwhile partnerships. This delicate balance raises questions about how we update our evaluations of otherwise upstanding people who have had a temporary moral lapse, particularly when these lapses occur within the context of dilemmas that force trade-offs between competing moral values. In this paper, we focus on answering the question of how personal moral values shape revisions of moral character attributions.

1.1. *The good true self hypothesis*

On the one hand, research on commonsense intuitions about the "true self" suggests that people will be most forgiving of violations against the moral categories they hold dear, because people tend to essentialize morally positive traits and believe that others' "true selves" are virtuous (Strohminger et al., 2017). For example, people typically think that somebody who acts in ways that foster racial equality is acting more in accordance with their true self than somebody who acts in racially discriminatory ways (Newman et al., 2015). Additionally, political conservatives think that other people are expressing their true selves when they become more patriotic, more religious, or more monogamous (reflecting values that are frequently endorsed by conservatives), whereas they think this is less true of people whose values become more aligned with those of political liberals; the inverse pattern is found for liberals (Newman et al., 2014). In other words, personal transformations are most likely to be described as reflecting the true self when they are aligned with perceivers' own values. These tendencies persist across several cultures and are found even in misanthropes (De Freitas et al., 2018). On the basis of this evidence, it has been concluded that "there is a consistent propensity to believe that each and every one of us possesses a good true self" (De Freitas et al., 2017, p. 636).

If we indeed hold a robust default expectation that others will share our own moral values in their deepest core, it seems that we should have a tendency to enduringly impute others with the moral characteristics that we value most. If we witness a typically kind person behaving cruelly in a particular situation, this theoretical perspective suggests that we should overlook the unkindness – perhaps attributing this unexpected immoral action to a situational or superficial feature (like exhaustion) – thus preserving a belief that even a person who is occasionally unkind remains kind in their deepest core. Furthermore, this tendency should be particularly evident amongst people who consider kindness to be a cornerstone of morality. In sum, the "Good True Self Hypothesis" leads to the prediction that observers will *absolve* moral lapses more readily for more highly valued moral categories, based on the expectation that others are fundamentally good.

1.2. *Negativity dominates trait inferences in the moral domain*

A separate body of research focused on the evolutionary costs of optimism would lead to the opposite prediction, namely that people will be especially unforgiving when judging others' moral character. From this perspective, in situations of uncertainty, false negatives are often asymmetrically more costly than false positives, so even a single lapse should cause observers to be wary. Cognition is reliably biased in certain contexts to avoid costly mistakes (Haselton & Nettle, 2006), causing people to preferentially attend to negative information about others (Baumeister et al., 2001; Rozin & Royzman, 2001). This asymmetry especially applies to the moral domain, due to the risks of exploitation in cooperative situations (Ybarra, 2002). Many researchers have converged on the conclusion that immoral behaviors are considered more diagnostic of character traits than morally positive behaviors (see Reeder & Brewer, 1979; Skowronski & Carlston, 1989). This perception in turn facilitates greater updating of social perceptions for immoral behaviors than for moral behaviors (Reeder & Coovert, 1986); people have a lower threshold for downgrading others' moral character after a misdeed compared to upgrading their moral character after a good deed (Klein & O'Brien, 2016). Indeed, people who engage in an isolated immoral action are frequently perceived as setting off on a downward moral trajectory that will lead them to become increasingly immoral (Anderson et al., in press). Thus, single immoral actions are often disproportionately weighted over patterns of continuous moral adherence for influencing moral judgments.

However, there may be important boundary conditions for negativity dominance in judgments of moral character. For example, immoral actions that typically yield a strong negativity bias (e.g., acting dishonestly) often do not result in dispositional inferences when they are performed for higher moral reasons, like saving another person's life (Brown et al., 2005). This research suggests that people may be resistant to updating their moral character attributions if others are acting in ways that can be construed as morally good. Additionally, the literature on negativity dominance has focused on aggregate effects, rather than on investigating individual differences in evaluations of moral character, and so little is known about how observers' particular moral commitments impact tendencies to readily update attributions of moral character.

1.3. *The moral stringency hypothesis*

The current studies move beyond examining overall patterns in how people update their attributions of moral character, in order to investigate the influence of observers' particular moral value hierarchies. Specifically, we test the possibility that people might be *especially* prone to place disproportionate weight on violations of the moral values they prioritize most. This prediction can be derived from the same error-management perspective that explains the negativity biases reviewed above. Given that choices about affiliation are most consequential in domains that are most meaningful, it is reasonable to expect that tendencies toward being unforgiving will be exacerbated when monitoring whether people possess prioritized moral values. We therefore propose the "Moral Stringency Hypothesis": the hypothesis that others should quickly lose membership in people's most highly valued moral categories after a single deviation. Like the Good True Self Hypothesis, this hypothesis predicts that moral categorization should be impacted by observers' own moral values, but it generates the opposite prediction for how this impact can be expected to occur.

Two papers (to our knowledge) have yielded findings that are broadly consistent with the Moral Stringency Hypothesis, by demonstrating that negative actions considered highly immoral by observers are perceived to be particularly diagnostic of underlying traits, such that negative attributions vary alongside the strength of moral commitments. For example, political conservatives deem burning an American flag to be more indicative of being an unpatriotic person as compared to political liberals, whereas political liberals are more likely to perceive shouting homophobic slurs as indicative of being a homophobic person (Meindl et al., 2016). Similarly, people who value "binding values" (i.e., loyalty, respect, and purity) more than "individualizing values" (i.e., care and fairness) are more likely to think violations of binding values are caused by dispositional traits rather than situational factors, whereas people who primarily value individualizing values are more likely to think violations of care and fairness are caused by dispositions (Niemi et al., 2023). These papers both demonstrate that moral convictions can lead observers to become more likely to infer that immoral actions are diagnostic of bad character traits. Thus, the more people prioritize particular moral values, the more they should infer that others do not possess the corresponding moral traits whenever they detect a lapse. In other words, beliefs that other people possess morally virtuous essences are likely to endure only when these people prioritize observers' deeply held values across all situations.

1.4. Overview of studies

In sum, the Good True Self Hypothesis predicts that people should reliably categorize others as belonging to their most cherished moral categories, even after a lapse, since believing in a fundamentally good "true self" entails reliably projecting personal moral values onto others' deepest essences. Thus, positive relationships should exist between a person's relative prioritization of two competing moral values and their tendencies to judge other people as inherently possessing the most prioritized value. For example, the more somebody weights the moral value of loyalty over the moral value of fairness, the more enduringly they should consider somebody to be a loyal person (and the less enduringly they should consider somebody to be a fair person) after a single lapse. The Moral Stringency Hypothesis, by contrast, predicts that people will be quickest to characterize others as no longer belonging to their most cherished moral categories after a lapse. Thus, the more that somebody prioritizes one moral value over another, the *less* they should classify somebody as possessing the more highly cherished moral value if they witness any deviations. For example, the more somebody weights the moral value of loyalty over the moral value of fairness, the more likely they should be to stop considering somebody to be a loyal person (and the less likely they should be to stop considering somebody to be a fair person) if this person ever prioritizes fairness over loyalty. In the current paper, we test these opposing predictions regarding the tenacity or tenuousness of people's judgments about others' membership in moral categories. This work expands on the existing literature in several notable ways.

First, our studies directly pit the Good True Self Hypothesis against the Moral Stringency Hypothesis. This is important because prior investigations into the good true self and prior investigations into negativity dominance in moral character attributions have generally proceeded independently from one another, despite focusing on similar phenomena. Although the evidence from the two research programs appears contradictory, there may be ways to reconcile them. For example, dispositional attributions indicating negativity biases could have been made about superficial aspects of others' character. To ensure alignment with the focus on others' "inner cores" from the good true self literature, we probed participants' judgments about others' deep essences in each of our studies.

Second, we examined participants' evaluations of people who faced difficult moral dilemmas and who made decisions that deviated from a typically upheld moral value in order to temporarily prioritize an alternative moral value. These sorts of "lapses" that involve sacrificing one positive value for another competing value are commonly seen in everyday situations. Furthermore, decisions in these cases are likely to be motivated by truly moral motives, rather than self-interested, callous motives or other tendencies that are antithetical to being a good person. In this way, we diverge from past research on negativity dominance, which has typically focused on the role of extremely immoral actions in shaping evaluations of moral character—for example, "putting razor blades in children's apples on Halloween" (Birnbaum, 1973) and "stealing money from a charity fund" (Reeder & Coovert, 1986). These kinds of psychopathic actions are clearly diagnostic of a rotten moral character, or at least clearly antisocial motivations, but are fortunately quite rare.

Third, by investigating how people think about resolutions of tradeoffs between particular values like helpfulness or impartiality, we were able to narrow in on how specific moral values are projected onto others, given meaningful individual differences in how different moral values are prioritized (see Graham et al., 2013). This focus moves beyond global assessments of morality (e.g., whether somebody is a good person), which has been the focus of other research programs on attributions of moral character.

Fourth, we directly investigate whether changes in evaluations of moral character are driven by evaluations of cooperative potential. The Moral Stringency Hypothesis rests on the observation that the alignment of moral priorities is crucial for successful social partnerships. Thus, character attributions may be driven by people's assessments that others who deviate from their own values are poor social partners—particularly when these others purport to share their values but then hypocritically violate them. In our second study, we test whether this potential explanatory mechanism might underlie variations in how people update their beliefs about others' moral character, in ways that move beyond existing research on partner choice and hypocrisy.

For each study, we report all measures, conditions, data exclusions, and sample size determinations. Hypotheses, methods, data collection procedures, exclusion criteria, and analyses were preregistered via the Open Science Framework (Study 1: https://osf.io/7qruv; Study 2: https://osf.io/az2ke). Additionally, we conducted a preliminary study (preregistered at https://osf.io/3kgnx) that we report in full in the Supplementary Materials. The Supplementary Materials, as well as the data and analysis code for all studies, are available at https://osf.io/4yfuw.

2. Study 1

How do people update their assessments of others' moral character (e.g., the extent to which somebody is a helpful person) upon learning that these others have lapsed in upholding a relevant value (e.g., after somebody sacrifices helpfulness for impartiality)? Additionally, how are these evaluative changes influenced by individual differences in moral values (e.g., variations in prioritizing helpfulness as compared to impartiality)? In Study 1, we addressed these questions by sequentially presenting participants with three pieces of information: an introduction of the target character, a description of a moral dilemma, and the character's resolution of the dilemma. Participants made judgments after each piece of information. This paradigm was inspired by related research on updating for moral blame (Monroe & Malle, 2019), predictions of future moral behavior (Lupfer et al., 2000), and impressions of moral character (Brambilla et al., 2019; Kim et al., 2020; Reeder & Coovert, 1986). In addition to evaluating moral character, participants were asked to rate their endorsement of various moral values. To increase the generalizability of our findings, we measured moral valuation in two ways: scenario-specific evaluations of particular moral actions and decontextualized ratings of abstract moral values.

The Moral Stringency Hypothesis predicts that decreased attributions of moral character after a lapse should become more pronounced as participants' prioritization of particular moral values increases. In contrast, the Good True Self Hypothesis predicts less updating in attributions of moral character as participants' prioritization of particular moral values increases.

2.1. Method

2.1.1. Participants

Participants were recruited via Prolific and paid $1.50 each. They were required to be United States residents and to have at least a 95% approval rating. People who had participated in our preliminary study (see the Supplementary Materials) were restricted from participating.

A power analysis indicated that a sample size of 118 participants would be needed to achieve 80% power for a medium effect size (d = .50). As each participant only responded to 1/4 of the vignettes, we multiplied this by 4, yielding a target sample size of 472 participants. In anticipation of needing to exclude up to 15% of our participants for missing attention checks, we aimed to test 550 participants. A total of 552 people completed the survey, of whom 97 missed at least one attention check—providing an inadequate open-ended description of one of the scenarios they read (n = 53) or missing more than one Winograd Schema question (n = 44). The final sample included 455 participants ($M_{\text{Age}} = 33.99$, $SD_{\text{Age}} = 11.59$; 241 women, 208 men, 6 non-binary; 74.07% White).[1]

2.1.2. Procedure

In the primary task, each participant evaluated five randomly selected scenarios from a total pool of 20. These scenarios involved realistic dilemmas pitting different moral values against one another. Nearly half of these dilemmas were inspired by research on Moral Foundations Theory (Graham et al., 2013), such as studies of trade-offs between fairness and loyalty (Waytz et al., 2013). For the remainder of the dilemmas, we drew from other literature exploring additional value trade-offs (e.g., Levine et al., 2020; Rottman et al., 2021). Specifically, we examined dilemmas involving care vs. loyalty, fairness vs. loyalty, care vs. obedience to authority, fairness vs. obedience to authority, frugality vs. generosity, devotion to family vs. diligence in one's work, impartiality vs. helpfulness, selflessness vs. purity, protectiveness vs. honesty, and humanitarianism vs. environmentalism. Each scenario was presented to participants in three sequential chunks. An example pitting honesty against protectiveness is as follows (for the full set of scenarios, see Table S2 in the Supplementary Materials):

> A nursing home coordinator, Julia, considers herself to be an honest person. Because of her strong moral commitment to telling the truth, Julia consistently refrains from lying. For instance, Julia is always upfront about difficult circumstances when talking with her patients.
>
> One day, a patient with memory loss asks Julia about his wife who, as Julia knows, died a long time ago. Julia knows that telling the patient that his wife has passed away will make him unhappy for hours before he again forgets about it due to his severe memory loss.
>
> Julia decides to lie to her patient by telling him that his wife went on a walk. Doing this made Julia uncomfortable, given her strong moral commitment to telling the truth, but she felt that this particular situation justified her action.

We obtained moral character judgments after the first paragraph and after the third paragraph, which allowed us to calculate the extent to which participants' evaluations changed after a target character failed to uphold a particular moral value in the context of a single dilemma. At both timepoints, participants were asked to indicate the extent to which each target character was a member of a specific moral category (e.g., "honest person") in the deepest, most essential core of his or her being, on a scale from −10 (not at all) to 10 (very much). This particular language (i.e., asking participants to consider "the deepest, most essential core" of others' beings) has been used in prior research on the true self (e.g., Newman et al., 2014) and so allowed for direct comparison with this previous work. After the second paragraph (when the dilemma was revealed, but before the character's action was revealed), participants were asked to judge how moral it would be for each character to resolve the dilemma in the way they eventually did, on a scale from −10 (morally wrong) to 10 (morally right). This provided a context-specific indicator of participants' moral values, which we used as our primary predictor variable.

We also explored two possible mechanisms of perceived moral character change by asking participants to make two additional judgments within each scenario. First, participants were asked to predict the extent to which the target character would behave consistently with the particular moral value in the future, on a scale from −10 (never) to 10 (always), both before and after the lapse was presented. Second, participants were asked to predict the percentage of people who, if placed in the target character's position, would act like the target character in deviating from the moral value under consideration, on a scale from 0% (nobody) to 100% (everybody). These questions provided additional insight into whether the diagnosticity (i.e., future informativeness and general prevalence) of characters' actions influenced participants' judgments about changes in moral character.

After the scenarios, participants rated the degree to which they valued each of the 16 abstract moral categories that were featured in the vignettes (i.e., being a caring person, being a loyal person, being a fair person, being an obedient person, being a pure person, being a selfless person, being a helpful person, being an impartial person, being a frugal person, being a generous person, being a devoted person, being a diligent person, being a protective person, being an honest person, being a humanitarian, and being an environmentalist). These were presented in a unique random order for each participant and were rated on a scale from 0 (not at all) to 10 (a great deal). This decontextualized measure served as an additional means of assessing participants' relative prioritization of different moral values.

Finally, we administered two individual differences measures of domain-general categorization tendencies, presented in random order. One was a previously validated 15-item Need for Closure Scale (Roets & Van Hiel, 2011), which measures the extent to which people dislike uncertainty and which is related to rigidity in categorization. Another was an exploratory "asymmetric mixtures" task, adapted from Noyes and Keil (2018), in which we asked participants to judge the extent to which 100 ounces of apple juice would still be apple juice if one ounce of urine was added to it, and the extent to which a person with primarily White ancestry would still be White if she discovered that she had one Black grandparent. This was meant to assess the amount that various participants prioritized distinctive elements when making non-moral categorization judgments. Finally, participants were asked demographic questions and debriefed.

2.2. Results

2.2.1. Computation of difference scores

Our primary dependent variable was moral character change, which was obtained by computing difference scores between participants' moral character judgments made prior to each lapse and their moral character judgments made after each lapse (such that more negative scores indicated a greater loss of moral category membership).[2] Similarly, we computed a difference score between predictions of future

[1] Although our sample of 455 participants did not reach our intended usable sample size of 472, a post-hoc power simulation indicated that we obtained 87% power or above for our observed effects.

[2] Of the 2275 judgments made by participants, 338 (14.86%) indicated a *gain* in moral category membership, rather than a loss. These responses were unexpected, so we did not preregister any decisions about how to handle them. In the end, we chose to retain these cases in our primary analyses, but the results remain consistent when instead excluding these judgments from all analyses (see the Supplementary Materials for analyses that exclude these responses, alongside additional exploratory analyses of these responses).

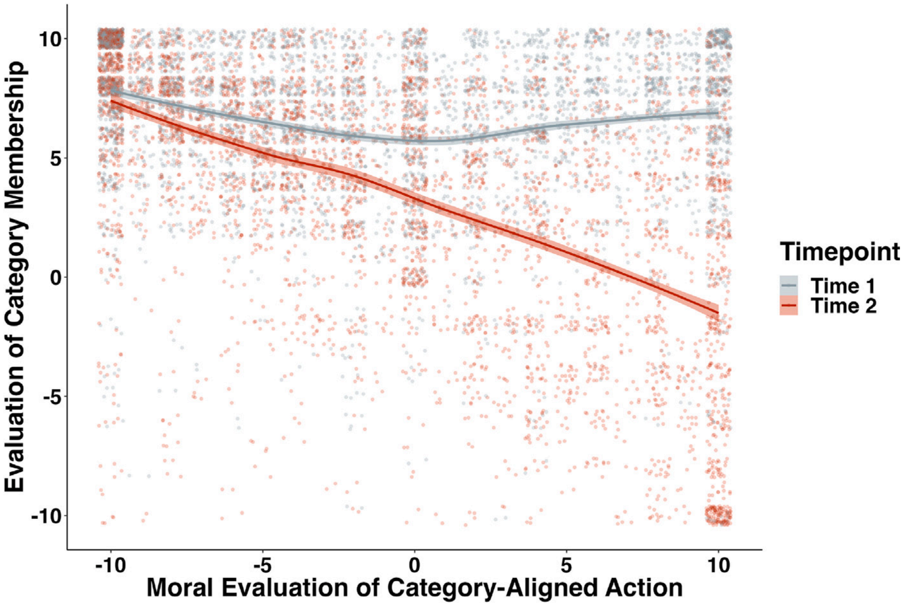

Fig. 1. Associations between category membership ratings at the two timepoints (before and after the moral lapse) and moral judgments for resolving specific moral dilemmas. Colored bands represent 95% CIs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

behavior made prior to each lapse and predictions of future behavior made after each lapse, which served as a measure of how much a moral lapse would lead participants to expect that the target character would no longer adhere to a particular moral category.[3]

2.2.2. *Primary analyses*

We conducted a series of linear mixed effects models predicting moral character change. In each, we accounted for the random intercepts of Scenario and Participant. By modeling these variables as random effects, we are able to generalize our results to a broader population of possible vignettes and participants, reducing concerns that our findings are an artifact of the particular scenarios that were created or the particular people who were tested (Judd et al., 2012).

First, we predicted perceived character change from participants' moral judgments of how the target characters should have acted within each scenario. Moral evaluations were coded such that higher values reflected judgments that the character should have acted in adherence with the value under consideration (and thus in alignment with the character's past behavior), while lower values reflected judgments that the character should have acted in accordance with the competing moral value (which is how the characters actually acted in the dilemmas). Moral evaluations were a strong predictor of perceived decreases in moral character, $b = -0.38$, $SE = 0.02$, $p < .001$. The more participants thought characters should adhere to the moral value in question, the greater the loss of moral category membership after a single lapse (see Fig. 1).

We then reran these analyses by substituting each scenario-specific moral evaluation with our more global, decontextualized measurement of moral values: Participants' relative prioritization of the two moral values at stake in each dilemma. This variable was coded such that higher values indicated greater prioritization of the moral value under consideration (i.e., the value underlying the target character's past behavior), and lower values indicated greater prioritization of the competing moral value (i.e., the value in line with the character's actual resolution of the dilemma). The relative moral valuations were also a strong predictor of the loss of category membership, $b = -0.21$, $SE = 0.04$, $p < .001$. Thus, the more participants valued the moral category being violated *in general*, relative to the competing moral category, the greater the loss of moral category membership after a lapse (see Fig. 2). Thus, our data provide consistent evidence supporting the Moral Stringency Hypothesis.

2.2.3. *Mechanisms underlying moral updating*

We also considered two possible mechanisms through which observers' moral values might shape their perceptions of others' moral character, and we tested whether these mechanisms supported alternative interpretations of our results. Specifically, one possible interpretation is that participants might expect behaviors they value to be more likely overall (Bear & Knobe, 2017), and that these beliefs about prevalence – rather than moral evaluations *per se* – could explain variation in the perceived loss of moral category membership across participants because participants weight lapses more heavily when they view them as more uncommon. Indeed, a simple correlation confirmed that the expected prevalence of category-violating behaviors was negatively correlated with scenario-specific moral evaluations, $r(2273) = -.44$, $p < .001$, meaning that participants' expectations for how other people *would* act in each scenario aligned with their judgments about how characters *should* act. Prevalence was uncorrelated with decontextualized moral valuation, however: $r(2273) = -.02$, $p = .377$. A second possible interpretation is that people who value a particular behavior may interpret momentary lapses as more indicative of future behavior, and that these predictions underlie judgments about moral category membership (Del Pinal & Reuter, 2017). Simple correlations confirmed that predictions of future behavior were negatively correlated with both scenario-specific moral evaluations, $r(2273) = -.47$, $p < .001$, and with decontextualized moral evaluations, $r(2273) = -.21$, $p < .001$.[4] These preliminary results both confirm past work and shed light on the

[3] A small number of difference scores (8.16%) indicated predictions that characters would act *more* in accordance with the category after the moral lapse. Again, this was unexpected. We retained these cases in our primary analyses, but report findings from analyses that excluded these responses in the Supplementary Materials; these exclusions do not change the findings.

[4] Although we did not preregister mediation analyses, these correlational results prompted us to explore whether estimates of the likelihood of lapse-consistent actions and estimates of future behavior each mediated the effects of moral evaluations on loss of category membership. Both mediation analyses demonstrated significant indirect effects, *ps* < .001 (see the Supplementary Materials for full results).

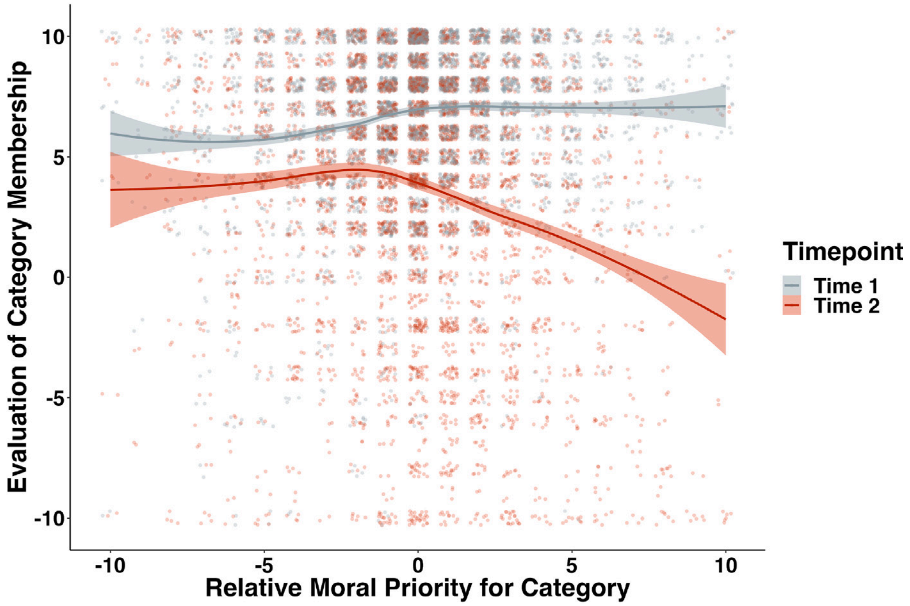

Fig. 2. Associations between category membership ratings at the two timepoints (before and after the moral lapse) and decontextualized ratings of moral values. Colored bands represent 95% CIs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mechanisms of moral judgments. However, our key prediction was that moral values shape assessments of moral character *beyond* their shared variance with descriptive beliefs about prevalence or future behavior.

To test this prediction, we ran an additional linear mixed effects model including these two other measured variables (i.e., change in predictions of future behavior and estimates of the prevalence of a lapse-consistent action) as well as both scenario-specific and decontextualized relative moral value scores. This full model confirmed our prediction: Each of the four predictor variables uniquely predicted decreases in attributions of moral character. Higher prevalence scores were associated with a reduced loss of category membership (b = −0.01, SE = 0.003, p = .002), meaning that participants were more likely to discount moral lapses that they thought were common. Additionally, expected changes in future behavior predicted changes in category membership (b = 0.81, SE = 0.02, p < .001), such that greater expected decreases in future behavior were associated with a greater loss of moral category membership. However, even when controlling for estimates of prevalence and future behavior, category judgments were *still* predicted by both scenario-specific moral evaluations (b = −0.12, SE = 0.01, p < .001) and decontextualized relative moral value scores (b = −0.08, SE = 0.02, p < .001), indicating that estimates of the likelihood of a lapse and of future behavior cannot fully account for variation in judgments about moral category membership.

Finally, we reran each of the previous three models, with the addition of all the individual differences measures—both of the two Asymmetric Mixtures questions (which were not strongly correlated with each other, $r(453)$ = .25, and thus not combined into a single index) and the Need for Closure measure (α = .86). Need for Closure weakly predicted reductions in category membership in the first model (with scenario-specific moral evaluations as a predictor), b = −0.37, SE = 0.17, p = .033, but otherwise these three variables did not significantly predict loss of category membership (all other ps > .09 in the first two models, all ps > .22 in the full model). These null results indicate that domain-general categorization tendencies are unlikely to account for the particular moral categorization tendencies that we observed in our study.

2.3. Discussion

Study 1 assessed how participants updated their categorization of others' "deepest, most essential core" after learning of single deviations from a broad range of moral values in the context of moral dilemmas. Initial categorization judgments were high regardless of participants' moral values, in line with previous research showing that people expect others who have previously acted morally to continue behaving in moral ways if they do not have any evidence to the contrary (Lupfer et al., 2000). Category membership dropped substantially following a moral lapse, despite target characters' stated track record of adhering to particular values. Although this finding is surprising from a Bayesian updating perspective, it aligns with prior research on the negativity bias (Klein & O'Brien, 2016; Lupfer et al., 2000) and the salience of morality in impression updating (Brambilla et al., 2019). Moving beyond previous work, the present research produced a novel insight: decreases in perceived moral character were most evident when participants themselves prioritized the moral values in question. This result held true both when priorities for particular moral values were measured by evaluations of how particular dilemmas should be resolved, and when these priorities were measured by relative differences in decontexualized ratings of moral values.

Participants' estimates of the general prevalence of violating particular moral values and their predictions of target characters' future behavior were each important predictors of the loss of moral category membership. Nonetheless, participants' decontextualized and scenario-specific moral values remained important predictors of moral categorization above and beyond these mechanisms, suggesting that differences in moral values *per se* importantly predict variations in attributions of moral character.

3. Study 2

Why did participants in Study 1 so readily diminish the character of others who deviated from their shared moral priorities? We considered it plausible that participants were particularly attuned to hypocritical actors who initially appeared to be good cooperative partners due to their mutually held values, but who later violated this assumption. Past research has found that people are especially harsh judges of others who hypocritically fail to uphold moral values that they have previously espoused, thus falsely advertising their quality as social affiliates (Effron & Monin, 2010; Jordan et al., 2017). This is due in part to perceptions that moral stances imply enduring commitments and carry expectations of behavioral consistency (Kreps & Monin, 2014). Since deviations from particular moral commitments can signal a general lack of moral integrity and trustworthiness (Kreps et al., 2017), it would be adaptive for people to have a hair-trigger reaction to signs of hypocrisy, particularly in cases where others violate their most cherished values.

If this interpretation is correct, then a very different pattern of evaluations should emerge for characters who do not lapse from their typical moral priorities, but instead remain consistent with their own moral compasses. Compared to people who temporarily deviate from both their typical moral values and participants' own values, people who consistently flaunt participants' values should not be considered to have more degraded moral character after they resolve a moral dilemma in a way that participants think is wrong. Indeed, if target characters demonstrate a steadfast commitment to alternative moral values, this may in fact increase evaluations of the degree to which they are good cooperative partners, even if moral character judgments remain similar.

On the other hand, moral character evaluations may instead primarily serve to help us track and cooperate with others who act in accordance with our own moral values, regardless of whether they express commitments to these values. If this is the case, then a target character who acts against a participant's moral values should decrease in moral category membership regardless of whether or not that action is aligned with the target character's own typical moral values. This alternative hypothesis therefore predicts that whether or not an action is a lapse will have less impact than whether or not it diverges from a participant's moral values.

The design and results of Study 1 do not allow us to clearly adjudicate between these two accounts of our findings. In Study 2, we tested these possible explanatory hypotheses by adding a manipulation of whether or not the target characters typically prioritize the moral values that are violated in the context of a dilemma. We additionally measured participants' assessments of the characters' cooperative potential before and after the characters resolved the dilemmas, to shed light on the possibility that moral character attributions are shaped by assessments related to partner choice. Overall, our aims in Study 2 were to conceptually replicate our previous study and to examine a potential mechanism for the moral stringency effect.

3.1. Method

3.1.1. Participants

Participants were recruited via Prolific and paid $3.00 each. They were required to be United States residents and to have at least a 95% approval rating. Individuals who had participated in the preliminary study or in Study 1 were restricted from participating.

Given the larger number of scenarios in this study and the counterbalanced design we employed (see below), a power analysis indicated that a sample size of 26 participants would be needed to achieve 80% power for a medium effect size (d = .50). Because each participant only responded to 1/16 of the vignettes, we multiplied this number by 16, yielding a target sample size of 416 participants. In anticipation of needing to exclude up to 20% of our participants (given the high exclusion rate in Study 1), and in order to ensure a well-powered dataset that was approximately the same size as the previous study, we aimed to test 550 participants. A total of 553 people completed the survey, of whom 40 missed at least one attention check—providing an inadequate open-ended description of one of the scenarios they read (n = 30) or missing more than one Winograd Schema question (n = 10).[5] Additionally, two participants had missing data for the primary questions of interest and so were excluded. The final sample included 511 participants (M_{Age} = 40.25, SD_{Age} = 13.33; 218 women, 284 men, 9 non-binary; 72.60% White).

3.1.2. Procedure

Each participant was presented with a set of four vignettes, out of a total of 64. There were 16 total sets, constructed by using a Latin Squares design. All participants saw two vignettes that involved lapses from prioritized moral values and two that did not.

In the Lapse condition, the vignettes were very closely matched (and in some cases nearly identical) to those in Study 1. Participants first read a general description of the target character, then read about a moral dilemma the character faced, and finally read about the character's decision to resolve the dilemma by acting against the moral value in question. For instance, a scenario pitting respectfulness against tolerance is as follows:

> Sarah, an accountant and a mother of three teenagers, has a lot of strong moral commitments. She consistently prioritizes being considerate and showing deference, and she considers being respectful to be the moral value that is of utmost importance to her.
>
> Recently, Sarah and her teenage children were invited to an Orthodox Jewish wedding. They were asked to wear clothing that fully covered their elbows and legs and hair, and they were told that she needed to sit with other women, separately from the men in attendance. Sarah's 16-year-old daughter, a staunch feminist, told Sarah that she did not feel comfortable abiding by these restrictions and planned to quietly disobey them. Sarah realized that she faced a moral dilemma: Should she be respectful of the Jewish traditions and require her daughter to adhere to the religious guidelines, or should she be tolerant of her daughter's wish to choose how she wants to dress and act?
>
> Sarah decided to allow her daughter to dress and sit however she wanted at the wedding. Even though being respectful is Sarah's primary moral value, she felt that this particular situation meant that she needed to act tolerantly instead of adhering to the religious guidelines.

In the No Lapse condition, there were no differences to the descriptions of the dilemmas or the characters' resolutions of the dilemmas. Instead, the only difference was that the characters were not described as being committed to the moral value in question. For the scenario above, Sarah was instead introduced as follows:

> Sarah, an accountant and a mother of three teenagers, has a lot of strong moral commitments. She consistently prioritizes being a good person. However, she does not consider being respectful to be one of her primary moral values.

Additionally, the target character's decision was described as being consistent with their values, and therefore (despite being the same decision made in the Lapse condition) not indicative of a moral lapse. For Sarah, this text was as follows:

> Sarah decided to allow her daughter to dress and sit however she wanted at the wedding. Especially because being respectful is not one of Sarah's primary moral values, she felt that this particular situation meant that she needed to act tolerantly instead of adhering to the religious guidelines.

Beyond introducing this experimental manipulation, we made a number of improvements to the vignettes in this study. Most notably, rather than creating separate scenarios for each inverse trade-off (e.g., having one scenario for sacrificing fairness for loyalty and a separate scenario for sacrificing loyalty for fairness), we created matched pairs of each scenario, such that some participants read about one value being sacrificed and other participants read about the competing value being sacrificed within the same context. Additionally, to further increase the generalizability of our findings, we broadened our set of dilemmas by decreasing the number of values drawn from Moral

[5] Because our previous final sample was smaller than we had anticipated, due to a greater-than-expected exclusion rate for missing our Winograd Schema "bot check" questions, we replaced the two most difficult Winograd Schema questions with two new questions that we expected would be easier to answer. Our intuitions were borne out.

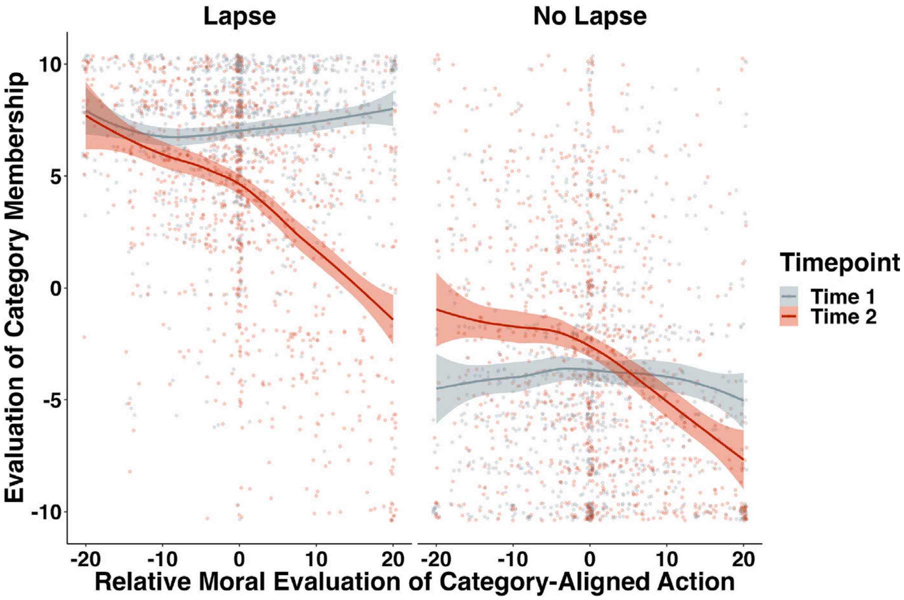

Fig. 3. Associations between category membership ratings at the two timepoints (before and after the dilemma) and moral judgments for resolving specific moral dilemmas, across the Lapse and No Lapse conditions. Colored bands represent 95% CIs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Foundations Theory and adding additional values inspired by a more extensive number of theoretical perspectives. For example, we included a dilemma involving the trade-off between merit and equality that has featured heavily in work on moral development (Baumard et al., 2012; Piaget, 1932), as well as including traits such as courageousness that are commonly discussed by virtue ethicists and that feature in newer theoretical frameworks of moral psychology (e.g., Curry et al., 2019). We also added a variant of the classic Heinz dilemma (Kohlberg, 1963) and a version of the classic lifeboat dilemma used to pit deontology against consequentialism (Greene et al., 2001). In the process of adding these additional dilemmas, we removed several trade-offs that had produced lopsided or extreme judgments in Study 1 (see Figs. S2 and S3 in the Supplementary Materials; for indications that the judgments became more balanced and moderate in Study 2, see Figs. S4 and S5 in the Supplementary Materials). In the end, we produced a set of vignettes in which characters had to choose between being honest or kind, merciful or obedient, frugal or generous, principled or humane, meritocratic or equitable, reliable or compassionate, religiously faithful or cooperative, caring or law-abiding, devoted to family or diligent in one's work, impartial or helpful, concerned for humans or concerned for the environment, loyal or altruistic, courageous or protective, respectful or tolerant, consequentialist or pacifist, and forgiving or fair. The full text of all scenarios is included in the Supplementary Materials (see Table S3).

Once again, participants made separate evaluations of moral character after reading the first and third paragraphs of the vignettes, with questions worded in the same way as in Study 1 (e.g., "Do you think that Sarah is a *respectful person* in the deepest, most essential core of her being?"). Following these evaluations of moral character, at both timepoints, we asked five new questions to assess participants' overall assessment of the target character's potential as a cooperative partner. These questions were:

1. "Would you consider [name] to have strong moral integrity?"
2. "Do you think that [name] is trustworthy?"
3. "How interested would you be in pursuing a friendship or business partnership with [name]?"
4. "How warm do you feel toward [name]?"
5. "Do you expect [name] to act in [respectful, humane, etc.] ways in the future?"

We additionally modified how participants made moral evaluations of the actions in the vignettes. In Study 1, participants were only asked to rate the extent to which it would be morally right for the target character to act in a way that went against the character's typical value (i.e., the way they ended up acting, which led to a lapse). In order to get a better sense of the extent to which participants valued the inconsistent way of acting as compared to the consistent way of acting, we asked participants to provide moral judgments of both possible actions in Study 2. Finally, we modified our decontextualized measure of abstract moral values such that participants explicitly rated the extent to which they prioritized one value over another, rather than providing separate endorsements of each value as participants had done in Study 1.

3.2. Results

3.2.1. Computation of difference scores

We again computed difference scores for each of our variables of interest. Moral character change was coded as in Study 1, with more negative scores indicating greater reductions in moral category membership at the second timepoint.

Difference scores were also computed for the questions assessing participants' feelings of warmth toward the target characters and their evaluations of the characters' moral integrity, trustworthiness, desirability as cooperative partners, and future likelihood to behave in accordance with the focal moral values. These scores were coded such that more negative scores indicated greater reductions at the second timepoint. We combined these variables into a single index indicating overall change in potential partner quality (α = .91).[6]

[6] Our preregistration indicated that we would create indices based on the results of exploratory factor analysis. A parallel analysis yielded a two-factor solution, with integrity, trustworthiness, and future behavior evaluations loading onto one factor, and partner desirability and warmth loading onto a second factor. However, these two factors were very highly correlated, r = .81, as were each of the five difference scores entered into the factor analysis, rs > .55. Based on these considerations (as well as there being only one eigenvalue above 1), we deemed it best to collapse all five items into a single index before proceeding with further analyses.

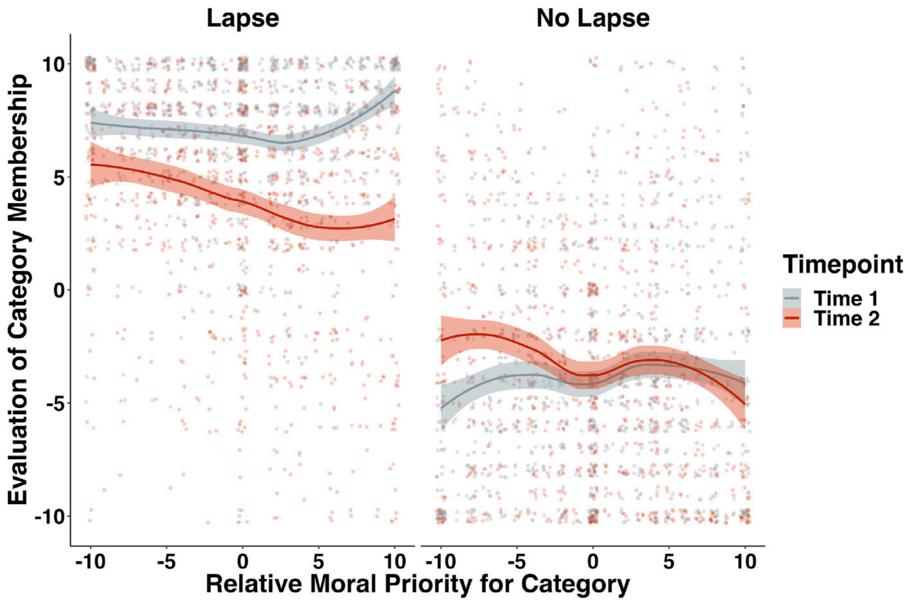

Fig. 4. Associations between category membership ratings at the two timepoints (before and after the dilemma) and decontextualized ratings of moral values, across the Lapse and No Lapse conditions. Colored bands represent 95% CIs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2.2. Primary analyses

We conducted a series of linear mixed effects models predicting moral character change, controlling for the random intercepts of Scenario and Participant, as well as random slopes when Condition was included in the models. First, we predicted perceived character change from participants' moral judgments of how the target characters should act within each scenario. Moral evaluations were coded such that higher values reflected judgments that the character should act in adherence with the focal moral value, while lower values reflected judgments that the character should violate the moral value in question (i.e., how the characters actually acted in the dilemmas). When examining results separately for the Lapse and No Lapse conditions, moral evaluations were a strong predictor of perceived changes in moral character, both for the Lapse condition, $b = -0.25$, $SE = 0.02$, $p < .001$ (replicating the findings from Study 1), and for the No Lapse condition, $b = -0.14$, $SE = 0.02$, $p < .001$. The more participants thought characters should adhere to the moral values in question, the greater the loss of moral category membership after a deviation from these values—and this was true regardless of whether or not the target characters were violating their own deeply held moral values (see Fig. 3).

Combining the data from the two conditions and modeling Condition as an additional fixed effect indicated that moral updating was more pronounced when target characters deviated from the focal moral values; perceived character change was significantly greater in the Lapse condition, $b = 3.80$, $SE = 0.30$, $p < .001$. There was also a significant interaction effect, such that perceived character change was most pronounced when participants thought it was morally best for characters to act consistently in the Lapse condition, $b = 0.10$, $SE = 0.02$, $p < .001$. This interaction effect provides some support for the hypothesis that perceptions of hypocrisy drive down moral category membership. However, it is notable that the interaction effect was relatively weak compared to the overall effect of participants' moral values, and that participants' moral values predicted perceived character change even when the actors did not act hypocritically.

We reran these analyses by substituting each scenario-specific moral evaluation with our more global, decontextualized measurement of moral values: Participants' relative prioritization of the two moral values at stake in each dilemma. This variable was coded such that higher scores indicated greater prioritization of the moral value in question, and lower scores indicated greater prioritization of the moral value that conflicted with the moral value in question (i.e., the value aligned with the character's actual resolution of the dilemma). Relative moral valuation was a strong predictor of decreases in category membership both for the Lapse condition, $b = -0.20$, $SE = 0.03$, $p < .001$ (again replicating Study 1), and for the No Lapse condition, $b = -0.15$, $SE = 0.03$, $p < .001$. Thus, the more participants prioritized the moral value being violated *in general* relative to the competing moral value, the greater the loss of moral category membership after the target character sacrificed this value. However, in the No Lapse condition, this was largely driven by valuations of the competing moral value driving up evaluations of moral character at the second timepoint (see Fig. 4).

Combining the data from the two conditions and modeling Condition as a fixed effect indicated that updating was most pronounced when target characters lapsed, with character change being significantly greater in the Lapse condition, $b = 3.89$, $SE = 0.28$, $p < .001$. However, there was no interaction effect between Condition and relative moral valuation, $b = 0.04$, $SE = 0.04$, $p = .258$, indicating that shared values, rather than perceptions of hypocrisy, are particularly influential for moral character evaluations.

Overall, the negative association between moral valuation and attributions of moral character dovetails with Study 1 by providing strong evidence in support of the Moral Stringency Hypothesis and against the Good True Self Hypothesis.

3.2.3. Exploration of explanatory mechanisms

To further examine the mechanism underlying the influence of participants' moral priorities on changes in evaluations of moral character, we added the composite variable indicating changes in ratings of the target characters' potential quality as cooperative social partners into each of the six models above. In all analyses, this variable assessing changes in partner quality was a significant predictor of changes in perceived moral character, $bs > 0.69$, $ps < .001$, such that greater decreases in positive assessments of the target character's potential as a good cooperative partner positively predicted greater decreases in their moral category membership, both in the Lapse and the No Lapse conditions. In four of the six models, participants' moral priorities became a non-significant predictor, suggesting that perceptions of moral character may strongly mediate the relationship between moral values and moral category membership. To test this directly, we ran several

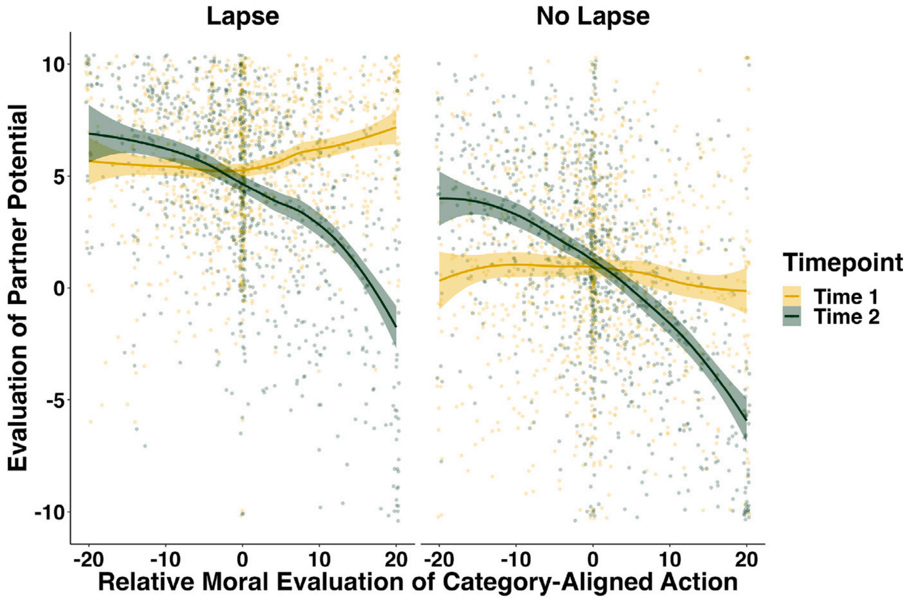

Fig. 5. Associations between participants' relative moral evaluations and their assessments of the target character's potential for being a good cooperative partner at the two timepoints (before and after the moral lapse), across the Lapse and No Lapse conditions. Colored bands represent 95% CIs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

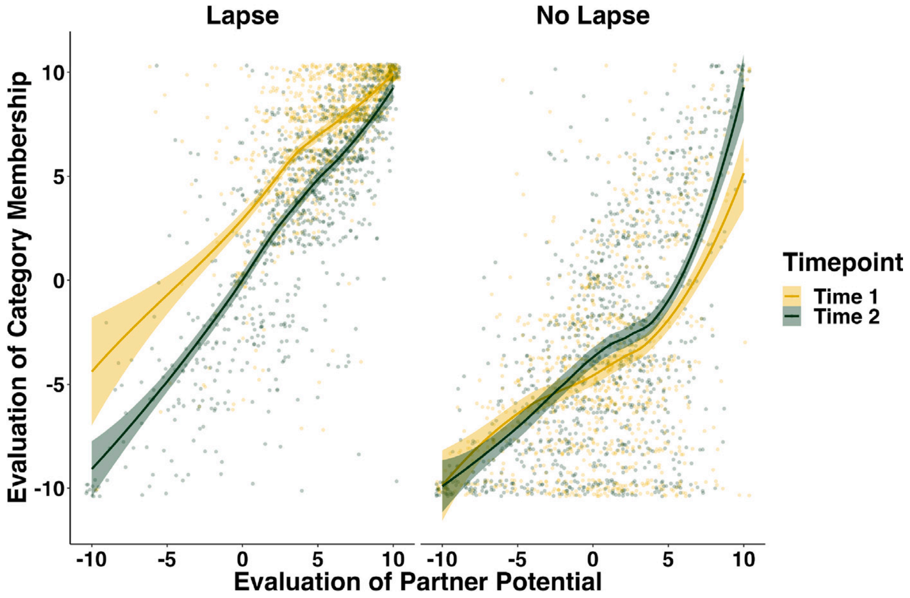

Fig. 6. Associations between category membership ratings at the two timepoints (before and after the moral lapse) and assessments of the target character's potential for being a good cooperative partner, across the Lapse and No Lapse conditions. Colored bands represent 95% CIs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

non-preregistered analyses. Additionally, we provide visualizations of the relationships between these variables in Figs. 5 and 6.

First, exploratory mixed effects models indicated that relative moral values strongly predicted changes in perceptions of partner quality, as did Condition. Perceptions that the target characters became worse potential partners at the second timepoint were increased by participants' judgments that the target characters should have resolved the dilemmas in ways that were consistent with the focal moral values, $b = -0.24$, $SE = 0.01$, $p < .001$, and by participants' relative prioritization of the focal abstract values, $b = -0.24$, $SE = 0.02$, $p < .001$. Additionally, lapses predicted greater overall change in partner quality than non-lapses, both in the model with specific moral evaluations, $b = 1.34$, $SE = 0.22$, $p < .001$, and in the model with abstract values, $b = 1.41$, $SE = 0.24$, $p < .001$. However, there was not a clear interaction between these variables (first model: $b = 0.03$, $SE = 0.02$, $p = .048$; second model: $b = 0.05$, $SE = 0.03$, $p = .080$). Furthermore, an additional mixed effects model found that changes in perceptions of partner quality strongly predicted changes in category membership attributions, $b = 0.93$, $SE = 0.03$, $p < .001$, as did the experimental condition, $b = 2.58$, $SE = 0.21$, $p < .001$. In this model, there was a significant interaction between these two variables, $b = -0.24$, $SE = 0.05$, $p < .001$. Thus, while perceptions of partner quality tend to be strongly linked to changes in moral category membership, this association becomes stronger when target characters hypocritically deviate from their primary moral values.

Given these results, we additionally conducted exploratory mediation analyses (accounting for random intercepts for participants). These indicated that there was a significant indirect effect of the target character's perceived partner quality, both in the Lapse and No Lapse conditions. In the Lapse condition, when using scenario-specific moral evaluations as the predictor variable, the indirect effect ($b = -0.22$, $p < .001$) reduced the total effect of −0.25 to a direct effect of −0.03. When using global moral value priorities as the predictor variable, the indirect effect ($b = -0.18$, $p < .001$) reduced the total effect of −0.15 to a direct effect of 0.03. In the No Lapse condition, when using scenario-specific moral evaluations as the predictor variable, the indirect effect ($b = -0.11$, $p < .001$) reduced the total effect of −0.12 to a direct effect of −0.01. When using global moral value priorities as the predictor variable, the indirect effect ($b = -0.12$, $p < .001$) reduced the total effect of −0.15 to a direct effect of −0.02.

3.3. Discussion

Study 2 provided additional support for the Moral Stringency Hypothesis. As in Study 1, stronger commitments to particular moral values predicted greater reductions in attributions of moral character. This was particularly clear when moral priorities were assessed by evaluations of the specific dilemmas that target characters faced, as compared to ratings of prioritization for abstract moral virtues. This suggests that participants may be less concerned with others' consistent adherence to general moral values and more concerned about others acting in ways that they endorse within particular situations, such that their moral categorization judgments might reflect more nuanced and context-specific assessments of whether they consider somebody to be a good person.

This interpretation is supported by insights from the additional experimental manipulation included in Study 2. This manipulation allowed us to obtain ratings of the same characters engaged in the same actions, with one slight difference: In the Lapse condition, a particular moral value was described as the target character's utmost moral concern, and in the No Lapse condition, this particular moral value was described as not being among the target character's most prioritized moral concerns. As expected, there was a large overall effect of the experimental condition; target characters who prioritized a particular moral value were much more likely to be evaluated as belonging within the corresponding moral category as compared to target characters who did not prioritize the moral value. However, the role of participants' own moral values was similar across these two conditions. Regardless of whether the target characters acted hypocritically by violating their prioritized moral value in a dilemma, or instead acted consistently by continuing to downplay the moral value at stake, these target characters were considered to decline in their moral character by participants who highly prioritized the moral values at stake. Additionally, this effect was explained by reductions in perceptions of the target character's potential to be a good cooperative partner to a similar extent across the Lapse and No Lapse conditions. Even though participants' higher valuations of particular moral commitments seemed to yield greater perceptions of a weak conscience amongst target characters who deviated from their typical moral priorities, perceptions of flip-flopping or poor adherence to values cannot fully explain the changes in moral character assessments that we observed. Instead, people appear to primarily update their perceptions of moral character based on their own cherished moral values, rather than updating their perceptions based on an interaction between their own values and targets' values.

Our results are especially noteworthy in light of past research indicating that people are particularly concerned about others who do not practice what they preach or whose moral values change over time, regardless of whether these hypocritical actors shift toward or away from observers' own beliefs (Kreps et al., 2017). Even though hypocrites have the makings of particularly bad cooperative partners, we found that target characters' deviation from participants' values was more important in determining moral character evaluations than hypocrisy. This finding can perhaps help to shed light on findings that hypocrites often, but not always, generate greater blame than non-hypocritical wrongdoers (see Effron et al., 2018; Jordan & Sommers, 2022).

4. General discussion

We each strive to uphold a plurality of competing moral values, so it is often difficult (and perhaps undesirable) for us to stay true to a particular moral value in all circumstances (Graham et al., 2015). Nevertheless, oscillating between competing moral goals can lead to negative perceptions of moral character (Kreps et al., 2017). The present research investigated how people evaluate others who typically adhere to a specific moral value but who decide to prioritize a competing moral value when faced with a dilemma. Across two studies, our data indicated that observers perceive more diminished moral character in people who deviate from values they care about more deeply, yielding strong support for the Moral Stringency Hypothesis. Our tendency to readily eschew others from our most cherished moral categories may be adaptive, given that moral category labels are critical social signals that influence reputation and therefore partner choice. In support of this interpretation, Study 2 showed that changes in participants' evaluations of moral character were strongly predicted by changes in the characters' perceived potential as social partners. Thus, we may be particularly strict judges when other people deviate from moral principles we care most about because we view these violations as being especially threatening to our potential social interactions.

4.1. Contributions to existing literature

4.1.1. Negativity effects are driven by those who care most

Consistent with previous research on negativity dominance (e.g., Reeder & Coovert, 1986; Rozin & Royzman, 2001), participants were generally quick to update their moral character inferences after only a single deviation from a particular moral standard. Importantly, the current results move beyond the existing literature on negativity dominance by examining character attributions in the wake of difficult moral dilemmas involving conflicting values, rather than focusing on blatantly and extremely immoral actions.

We additionally add to the existing literature by showing how personal commitments to certain ideals can exacerbate the degree to which lapses impact categorization (for convergent evidence, see Meindl et al., 2016; Niemi et al., 2023). However, we see our findings as broadly consistent with prior work on negativity dominance. Our results also align with previous studies showing that the asymmetrical weight of negativity is less likely to be observed for actions that are less extreme, perhaps because they are perceived to be less diagnostic (see Rusconi et al., 2020). This suggests a possible explanation for our findings. If moral actions that violate an evaluator's most highly cherished moral values are thought to be relatively more extreme (Wojciszke et al., 1993), or if they trigger more negative emotions (Trafimow et al., 2005), they may be thought to be more indicative of character and in turn weighted more heavily than the actor's previous track record of morally positive actions, thus leading to greater updating.

Our research also furthers our understanding of negativity dominance by showing when it is most likely to be observed. Past research has found that some moral violations (in particular, "hierarchically restrictive" traits like dishonesty) are treated as more diagnostic than others (in particular, "partially restrictive" traits like uncooperativeness), such that the inference that somebody is honest is more readily disconfirmed by a single violation than the inference that somebody is cooperative (Trafimow et al., 2005; Trafimow & Trafimow, 1999). Although the best method for categorizing various forms of immorality as hierarchically or partially restrictive has not been fully clarified by past work, it is generally suggested that these divisions are dictated by widely held schemas linking behaviors to dispositions (Reeder & Brewer, 1979) or by the content of the actions themselves. For example, some researchers have suggested that hierarchically restrictive trait dimensions and partially restrictive trait dimensions might roughly map onto Kant's distinction between perfect and imperfect duties (Trafimow et al., 2005). Thus, it has not previously been considered that people's differing moral values might yield distinct senses of which values are hierarchically restrictive (and thus overriding) or partially restrictive (and thus allowing of exceptions). The studies presented here demonstrate that observers' own moral values may be more relevant for determining the extent of negativity dominance than the content of the moral values themselves.

4.1.2. Others' true selves are not always considered morally good

The present results are difficult to reconcile with the view that our most dearly held values are central to our perceptions of others' true selves and are believed to comprise a stable essence (Newman et al., 2014; Strohminger et al., 2017). Although the "Good True Self Hypothesis" is correct about observers' own moral values being powerful influences on their perception of others' character, the current findings did not provide any evidence that we enduringly conceptualize others as sharing our own values in their deepest core. Rather than readily discounting deviations from the moral values we prioritize most, these lapses are *particularly* likely to shape our evaluations of others.

However, as others have demonstrated (Lupfer et al., 2000), the negativity bias is primarily found in cases when people receive inconsistent information about others' character. Perhaps we indeed possess a default tendency to assume that others' true selves align with our own moral values, consistent with our pervasive tendency to trust others (Weiss et al., 2022), but this assumption is easily overridden by contrary evidence. If so, people might be inclined to project their values onto others in the absence of countervailing information, and before a deviation from a moral value occurs. For example, participants who strongly value loyalty might expect others to be more loyal overall, in the absence of any additional knowledge. In the current studies, exploratory analyses indicate that baseline evaluations of others (before a lapse) did not vary depending on participants' weighting of moral values, perhaps because we provided participants with such clear information about the target character's values. Nevertheless, further work is needed to determine why the current methodological approach provided results that seem to contradict the findings in the "true self" literature.

4.1.3. The influence of personal values on trust and partner choice

Our findings also add nuance to recent research indicating that people trust others who share their moral principles. For example, studies show that deontologists are more trusting of other deontologists while consequentialists are more trusting of other consequentialists (Bostyn et al., 2023). Somewhat paradoxically, however, our research shows that these tendencies may sometimes flip in cases where typically consequentialist or typically deontological individuals deviate from their modal tendencies. One possible interpretation based on the present findings is that people use shared moral values as cues to social partner quality, but they use other cues as well (e.g., hypocrisy and reliability) when contextually salient. In this way, people who highly value deontological approaches to morality may sometimes end up thinking more highly of a reliable consequentialist than an unreliable deontologist.

People who do not always act in accordance with the particular values they espouse are typically perceived as sending false signals, which can trigger negative social consequences and perceptions of hypocrisy (Jordan & Sommers, 2022). This typically leads to a strong preference for people whose moral commitments are overriding and not traded off even when other competing moral concerns are at play (Kreps et al., 2017). We predicted that individuals who temporarily lapse in upholding a moral value, even when their behavior is guided by other moral concerns, are likely to be discredited, while individuals who act similarly without having similar moral commitments would not. However, our results suggest that evaluations of moral character are similar for hypocrites and non-hypocrites. Instead, we may look for social partners who act in alignment with our moral values, whether or not they purport to share them.

4.1.4. Morality may be a unique dual character concept

This research additionally advances our understanding of social categories that exhibit a "dual character" (Knobe et al., 2013; Leslie, 2015), insofar as membership can be achieved either in a descriptive sense (acting in accordance with typical behaviors presupposed by the category) or in a normative sense (having values consistent with the category). Our research suggests that moral categories may be unlike many other dual character concepts, as neither the mere allegiance to a particular moral value nor tendencies to uphold a value most of the time are sufficient for category membership. Instead, our results suggest that being classified as belonging to a highly valued moral category requires both strong fidelity to the moral value and unyielding behavioral adherence. This may be because dual character concepts require a commitment to the category's values (Del Pinal & Reuter, 2017) and, in the case of moral categories, any behavioral lapse may indicate a lack of such commitment. Although category representativeness is often idealized (Bear & Knobe, 2017; Foster-Hanson & Lombrozo, 2022; Foster-Hanson & Rhodes, 2019), this may be particularly exacerbated in moral contexts.

4.2. Limitations and future directions

Our studies and others have indicated that people are extremely unforgiving in their ascriptions of moral character. Recent work also shows that people regard flexible moral stances with suspicion and prefer others who espouse absolutist moral stances that tolerate no exceptions (Huppert et al., 2023). Yet, given the ubiquity of trade-offs between competing moral values, as well as typical tendencies toward "moral mediocrity" (Schwitzgebel, 2019), it may be quite unusual for people to achieve and sustain unyielding commitments to particular moral values. It should therefore be exceptionally rare to consider others as embodying our most cherished moral values. However, it is possible that the stringent tendencies we observed in participants' attributions of moral character are not enduring. On the contrary, some research shows that impressions of immorality tend to be unstable, and that – despite updating their beliefs about immoral actors more readily than moral actors – observers are relatively quick to forgive agents who transgress (Siegel et al., 2018). Thus, future studies should examine later redemption and readmittance to moral categories. For example, although the current results indicate that people's moral values lead them to be especially stringent immediately after a moral lapse, it is also possible that these same values could lead people to be more lenient about readmittance to moral categories over time. This possibility, which could help to reconcile the two hypotheses tested here, would be an intriguing avenue for future investigation.

Future research should also investigate how changes in evaluations of moral character are impacted by single as opposed to repeated lapses. Previous work has indicated that, while consistent lapses in moral behavior are generally judged as indicative of hypocrisy, even single lapses can sometimes be judged as equally hypocritical as repeated lapses, like when a priest commits a single act of adultery (Alicke et al., 2013). Because people are often insensitive to frequency or scope of immoral actions (Hsee & Rottenstreich, 2004; Rottman & Young, 2019; Rozin & Royzman, 2001), it is likely that an initial moral lapse will be weighted much more heavily than subsequent lapses, but this should be directly explored.

We attempted to sample broadly from the moral domain in constructing our scenarios. Nonetheless, it would be informative to examine whether the effects we uncovered extend to other dilemmas, including those involving non-moral values. Because moral values are unique in yielding expectations of overriding commitments (Kreps et al., 2017) and because they are particularly likely to yield negativity biases in person perception (see Rusconi et al., 2020), we expect that non-moral lapses will result in less stringent character attributions. However, this is an open empirical question that will necessitate a clear differentiation between moral and non-moral values and that will require creativity in generating scenarios that pit non-moral values against one another. We anticipate that this will be difficult to accomplish, both because the boundaries of the moral domain are amorphous (Levine et al., 2021; Sinnott-Armstrong & Wheatley, 2014) and because we anticipate that it may be rare for dilemmatic situations to arise where people must choose between competing non-moral values in ways that seem threatening to their character. For instance, a person may highly value being creative and may also highly value being successful, but because neither of these values typically transcends situational factors, it would be surprising if an artist was no longer considered to be a creative person deep down if a particular context led them to decide to accomplish a non-creative ambition. We suspect that very few values are clearly non-moral and are nevertheless considered to have the kind of paramount status that moral values often have.

Another direction for future research would be to test whether our effects are modulated by a variety of factors that are relevant to partner choice, such as participants' relative social standing and the perceived availability of other morally upstanding partners (Barclay, 2013). Additional studies should examine whether our findings apply only to evaluations of strangers, when there may be ambiguous potential for establishing an affiliative relationship, or whether they also extend to judgments of others with whom we have existing personal relationships. Given recent research indicating that dispositional explanations of moral violations are primarily made for distant others (Niemi et al., 2023), we expect that our findings will hold mainly for unfamiliar others, and perhaps only for strangers who are not from disliked or stigmatized groups for whom partner choice is a less relevant motivation. Finally, because our research utilized convenience samples from the United States, replication in other populations will be critical to discern the generalizability of the present findings.

4.3. Conclusion

These studies provide the clearest evidence to date that others' moral character traits are not always perceived through rose-colored glasses, particularly when observers have strong convictions about specific moral values. Even when considering the deepest core of a person's being, classification into highly valued moral categories is strictly reserved for those who inexorably adhere to moral ideals. This occurs because even brief deviations from our own moral values are used as indications that others have plummeted in their potential to be social partners of high quality. Our tendency toward moral stringency carries profound implications for blame, forgiveness, and our notions about who others truly are, deep down. By gaining a better understanding of how we update our beliefs about others' moral character, we may be better able to challenge our assumptions about criminals, mend broken friendships, and gain a more optimistic view of our social worlds.

CRediT authorship contribution statement

Joshua Rottman: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft, Supervision, Funding acquisition. **Emily Foster-Hanson:** Conceptualization, Methodology, Writing – review & editing. **Sam Bellersen:** Methodology, Investigation.

Data availability

All data and analysis scripts are available on the Open Science Framework, at https://osf.io/4yfuw.

Acknowledgements

We are grateful to Nick Kroll, Walter Sinnott-Armstrong, and Chris Crandall for fruitful discussions; Casey Lewry and David Kinney for helpful comments on a previous draft of the paper; and the Franklin & Marshall College Committee on Grants for funding this research.

Appendix A. Supplementary data

Supplementary material can be found online at https://doi.org/10.1016/j.cognition.2023.105570 or https://osf.io/4yfuw.

References

Alicke, M., Gordon, E., & Rose, D. (2013). Hypocrisy: What counts? *Philosophical Psychology*, *26*(5), 673–701. http://dx.doi.org/10.1080/09515089.2012.677397.

Anderson, R. A., Ruisch, B. C., & Pizarro, D. A. On the highway to hell: Slippery slope perceptions in judgments of moral character. Personality and Social Psychology Bulletin, in press, http://dx.doi.org/10.1177/01461672221143022.

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behaviour*, *34*(3), 164–175. http://dx.doi.org/10.1016/j.evolhumbehav.2013.02.002.

Batson, C. D. (2016). *What's wrong with morality? A social-psychological perspective*. New York, NY: Oxford University Press.

Baumard, N., Mascaro, O., & Chevallier, C. (2012). Preschoolers are able to take merit into account when distributing goods. *Developmental Psychology*, *48*(2), 492–498. http://dx.doi.org/10.1037/a0026598.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. http://dx.doi.org/10.1037/1089-2680.5.4.323.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, *167*, 25–37. http://dx.doi.org/10.1016/j.cognition.2016.10.024.

Birnbaum, M. H. (1973). Morality judgment: Test of an averaging model with differential weights.. *Journal of Experimental Psychology*, *99*(3), 395–399. http://dx.doi.org/10.1037/h0035216.

Boehm, C. (2012). *Moral origins: The evolution of virtue, altruism, and shame*. New York, NY: Basic Books.

Bostyn, D. H., Chandrashekar, S. P., & Roets, A. (2023). Deontologists are not always trusted over utilitarians: revisiting inferences of trustworthiness from moral judgments. *Scientific Reports*, *13*(1), 1665. http://dx.doi.org/10.1038/s41598-023-27943-3.

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology*, *82*, 64–73. http://dx.doi.org/10.1016/j.jesp.2019.01.003.

Brambilla, M., & Leach, C. W. (2014). On the importance of being moral: The distinctive role of morality in social judgment. *Social Cognition*, *32*(4), 397–408. http://dx.doi.org/10.1521/soco.2014.32.4.397.

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. *Advances in Experimental Social Psychology*, *64*, 187–262. http://dx.doi.org/10.1016/bs.aesp.2021.03.001.

Brown, J., Trafimow, D., & Gregory, W. L. (2005). The generality of negative hierarchically restrictive behaviours. *British Journal of Social Psychology*, *44*(1), 3–13. http://dx.doi.org/10.1348/014466604X23455.

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47–69. http://dx.doi.org/10.1086/701478.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences*, *21*(9), 634–636. http://dx.doi.org/10.1016/j.tics.2017.05.009.

De Freitas, J., Sarkissian, H., Newman, G. E., Grossmann, I., De Brigard, F., Luco, A., & Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science*, *42*, 134–160. http://dx.doi.org/10.1111/cogs.12505.

Del Pinal, G., & Reuter, K. (2017). Dual character concepts in social cognition: Commitments and the normative dimension of conceptual representation. *Cognitive Science*, *41*, 477–501. http://dx.doi.org/10.1111/cogs.12456.

Effron, D. A., & Monin, B. (2010). Letting people off the hook: When do good deeds excuse transgressions? *Personality and Social Psychology Bulletin*, *36*(12), 1618–1634. http://dx.doi.org/10.1177/0146167210385922.

Effron, D. A., O'Connor, K., Leroy, H., & Lucas, B. J. (2018). From inconsistency to hypocrisy: When does "saying one thing but doing another" invite condemnation? *Research in Organizational Behavior*, *38*, 61–75. http://dx.doi.org/10.1016/j.riob.2018.10.003.

Flanagan, O. (2017). *The geography of morals: Varieties of moral possibility*. New York, NY: Oxford University Press.

Foster-Hanson, E., & Lombrozo, T. (2022). How "is" shapes "ought" for folk-biological concepts. *Cognitive Psychology*, *139*, Article 101507. http://dx.doi.org/10.1016/j.cogpsych.2022.101507.

Foster-Hanson, E., & Rhodes, M. (2019). Is the most representative skunk the average or the stinkiest? Developmental changes in representations of biological categories. *Cognitive Psychology*, *110*, 1–15. http://dx.doi.org/10.1016/j.cogpsych.2018.12.004.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. http://dx.doi.org/10.1037/a0034726.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *47*, 55–130. http://dx.doi.org/10.1016/B978-0-12-407236-7.00002-4.

Graham, J., Meindl, P., Koleva, S., Iyer, R., & Johnson, K. M. (2015). When values and behavior conflict: Moral pluralism and intrapersonal moral hypocrisy. *Social and Personality Psychology Compass*, *9*(3), 158–170. http://dx.doi.org/10.1111/spc3.12158.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. http://dx.doi.org/10.1126/science.1062872.

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, *10*(1), 47–66. http://dx.doi.org/10.1207/s15327957pspr1001_3.

Hsee, C. K., & Rottenstreich, Y. (2004). Music, pandas, and muggers: On the affective psychology of value. *Journal of Experimental Psychology: General*, *133*(1), 23–30. http://dx.doi.org/10.1037/0096-3445.133.1.23.

Huppert, E., Herzog, N., Landy, J. F., & Levine, E. (2023). On being honest about dishonesty: The social costs of taking nuanced (but realistic) moral stances. *Journal of Personality and Social Psychology*, *125*(2), 259–283. http://dx.doi.org/10.1037/pspa0000340.

Jordan, J. J., & Sommers, R. (2022). When does moral engagement risk triggering a hypocrite penalty? *Current Opinion in Psychology*, *47*, 101404. http://dx.doi.org/10.1016/j.copsyc.2022.101404.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science*, *28*(3), 356–368. http://dx.doi.org/10.1177/0956797616685771.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. http://dx.doi.org/10.1037/a0028347.

Kim, M., Park, B., & Young, L. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences*, *24*(2), 101–111. http://dx.doi.org/10.1016/j.tics.2019.12.001.

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. http://dx.doi.org/10.1521/soco.2016.34.2.149.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, *127*(2), 242–257. http://dx.doi.org/10.1016/j.cognition.2013.01.005.

Kohlberg, L. (1963). The development of children's orientations toward a moral order I. Sequence in the development of moral thought. *Vita Humana*, *6*, 11–33.

Kreps, T. A., Laurin, K., & Merritt, A. C. (2017). Hypocritical flip-flop, or courageous evolution? When leaders change their moral minds. *Journal of Personality and Social Psychology*, *113*(5), 730–752. http://dx.doi.org/10.1037/pspi0000103.

Kreps, T. A., & Monin, B. (2014). Core values versus common sense: Consequentialist views appear less rooted in morality. *Personality and Social Psychology Bulletin*, *40*(11), 1529–1542. http://dx.doi.org/10.1177/0146167214551154.

Leslie, S.-J. (2015). "Hillary Clinton is the only man in the Obama Administration": Dual character concepts, generics, and gender. *Analytic Philosophy*, *56*(2), 111–141. http://dx.doi.org/10.1111/phib.12063.

Levine, E. E., Roberts, A. R., & Cohen, T. R. (2020). Difficult conversations: Navigating the tension between honesty and benevolence. *Current Opinion in Psychology*, *31*, 38–43. http://dx.doi.org/10.1016/j.copsyc.2019.07.034.

Levine, S., Rottman, J., Davis, T., O'Neill, E., Stich, S., & Machery, E. (2021). Religious affiliation and conceptions of the moral domain. *Social Cognition*, *39*(1), 139–165. http://dx.doi.org/10.1521/soco.2021.39.1.139.

Lupfer, M. B., Weeks, M., & Dupuis, S. (2000). How pervasive is the negativity bias in judgments based on character appraisal? *Personality and Social Psychology Bulletin*, *26*(11), 1353–1366. http://dx.doi.org/10.1177/0146167200263004.

Meindl, P., Johnson, K. M., & Graham, J. (2016). The immoral assumption effect: Moralization drives negative trait attributions. *Personality and Social Psychology Bulletin*, *42*(4), 540–553. http://dx.doi.org/10.1177/0146167216636625.

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*, *116*(2), 215–236. http://dx.doi.org/10.1037/pspa0000137.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin*, *40*(2), 203–216. http://dx.doi.org/10.1177/0146167213508791.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, *39*(1), 96–125. http://dx.doi.org/10.1111/cogs.12134.

Niemi, L., Doris, J. M., & Graham, J. (2023). Who attributes what to whom? Moral values and relational context shape causal attribution to the person or the situation. *Cognition*, *232*, 105332. http://dx.doi.org/10.1016/j.cognition.2022.105332.

Noyes, A., & Keil, F. C. (2018). Asymmetric mixtures: Common conceptual priorities for social and chemical kinds. *Psychological Science*, *29*(7), 1094–1103. http://dx.doi.org/10.1177/0956797617753562.

Piaget, J. (1932). *The moral judgment of the child*. New York, NY: Harcourt.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*(1), 61–79. http://dx.doi.org/10.1037/0033-295X.86.1.61.

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition*, *4*(1), 1–17. http://dx.doi.org/10.1521/soco.1986.4.1.1.

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences*, *50*(1), 90–94. http://dx.doi.org/10.1016/j.paid.2010.09.004.

Rottman, J., Crimston, C. R., & Syropoulos, S. (2021). Tree-huggers versus human-lovers: Anthropomorphism and dehumanization predict valuing nature over outgroups. *Cognitive Science*, *45*(4), e12967. http://dx.doi.org/10.1111/cogs.12967.

Rottman, J., & Young, L. (2019). Specks of dirt and tons of pain: Dosage distinguishes impurity from harm. *Psychological Science*, *30*(8), 1151–1160. http://dx.doi.org/10.1177/0956797619855382.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. http://dx.doi.org/10.1207/S15327957PSPR0504_2.

Rusconi, P., Sacchi, S., Brambilla, M., Capellini, R., & Cherubini, P. (2020). Being honest and acting consistently: Boundary conditions of the negativity effect in the attribution of morality. *Social Cognition*, *38*(2), 146–178. http://dx.doi.org/10.1521/soco.2020.38.2.146.

Schwitzgebel, E. (2019). Aiming for moral mediocrity. *Res Philosophica*, *96*(3), 347–368. http://dx.doi.org/10.11612/resphil.1806.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science*, *24*(2), 125–130. http://dx.doi.org/10.1177/0963721414553264.

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. http://dx.doi.org/10.1038/s41562-018-0425-1.

Sinnott-Armstrong, W., & Wheatley, T. (2014). Are moral judgments unified? *Philosophical Psychology*, *27*(4), 451–474. http://dx.doi.org/10.1080/09515089.2012.736075.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. http://dx.doi.org/10.1037/0033-2909.105.1.131.

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, *12*(4), 551–560. http://dx.doi.org/10.1177/1745691616689495.

Trafimow, D., Bromgard, I. K., Finlay, K. A., & Ketelaar, T. (2005). The role of affect in determining the attributional weight of immoral behaviors. *Personality and Social Psychology Bulletin*, *31*(7), 935–948. http://dx.doi.org/10.1177/0146167204272179.

Trafimow, D., & Trafimow, S. (1999). Mapping perfect and imperfect duties onto hierarchically and partially restrictive trait dimensions. *Personality and Social Psychology Bulletin*, *25*(6), 687–697. http://dx.doi.org/10.1177/0146167299025006004.

Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science*, *18*(8), 689–690. http://dx.doi.org/10.1111/j.1467-9280.2007.01961.x.

van Leeuwen, F., Park, J. H., & Penton-Voak, I. S. (2012). Another fundamental social category? Spontaneous categorization of people who uphold or violate moral norms. *Journal of Experimental Social Psychology*, *48*(6), 1385–1388. http://dx.doi.org/10.1016/j.jesp.2012.06.004.

Walker, L. J., & Hennig, K. H. (2004). Differing conceptions of moral exemplarity: Just, brave, and caring. *Journal of Personality and Social Psychology*, *86*(4), 629–647. http://dx.doi.org/10.1037/0022-3514.86.4.629.

Waytz, A., Dungan, J., & Young, L. (2013). The whistleblower's dilemma and the fairness–loyalty tradeoff. *Journal of Experimental Social Psychology*, *49*(6), 1027–1033. http://dx.doi.org/10.1016/j.jesp.2013.07.002.

Weiss, A., Burgmer, P., & Hofmann, W. (2022). The experience of trust in everyday life. *Current Opinion in Psychology*, *44*, 245–251. http://dx.doi.org/10.1016/j.copsyc.2021.09.016.

Wojciszke, B., Brycz, H., & Borkenau, P. (1993). Effects of information content and evaluative extremity on positivity and negativity biases. *Journal of Personality and Social Psychology*, *64*(3), 327–335. http://dx.doi.org/10.1037/0022-3514.64.3.327.

Ybarra, O. (2002). Naive causal understanding of valenced behaviors and its implications for social information processing.. *Psychological Bulletin*, *128*(3), 421–441. http://dx.doi.org/10.1037/0033-2909.128.3.421.